THE UNIVERSITY *of* EDINBURGH

ANALYSIS OF ROUTINELY COLLECTED REPEATED PATIENT OUTCOMES

CHRISTIAN HOLM HANSEN

PH.D. – THE UNIVERSITY OF EDINBURGH - 2014

PREFACE

Declaration. I declare that this thesis has been composed by me and that the work presented is my own. As a member of a research group the clinical studies described in this thesis were conceived, designed and conducted in collaboration with colleagues and were funded as part of a Programme Grant from Cancer Research UK (CRUK), ref. C5547/A7375. The study presented in Chapter 4 has now been published. The published article is included as an appendix. The work presented has not been submitted for any other degree or professional qualification.

Acknowledgements. I wish to thank Professors Gordon D Murray and Michael Sharpe for their time and excellent supervision, helpful advice, and for offering me the opportunity and time to undertake this work. Thanks are also due to the thousands of patients who provided the data behind much of the work presented here and who allowed these to be used for research. Finally to Amy, my 'PhD widow' of the past four years: thank you for putting up with me, I have you to thank for this.

Christian Holm Hansen
Edinburgh
November 2013

ABSTRACT

Clinical practice should be based on the best available evidence. Ideally such evidence is obtained through rigorously conducted, purpose-designed clinical studies such as randomised controlled trials and prospective cohort studies. However gathering information in this way requires a massive effort, can be prohibitively expensive, is time consuming, and may not always be ethical or practicable. When answers are needed urgently and purpose-designed prospective studies are not feasible, retrospective healthcare data may offer the best evidence there is. But can we rely on analysis with such data to give us meaningful answers?

The current thesis studies this question through analysis with repeated psychological symptom screening data that were routinely collected from over 20,000 outpatients who attended selected oncology clinics in Scotland. Linked to patients' oncology records these data offer a unique opportunity to study the progress of distress symptoms on an unprecedented scale in this population. However, the limitations to such routinely collected observational healthcare data are many. We approach the analysis within a missing data context and develop a Bayesian model in WinBUGS to estimate the posterior predictive distribution for the incomplete longitudinal response and covariate data under both Missing At Random and Missing Not At Random mechanisms and use this model to generate multiply imputed datasets for further frequentist analysis.

Additional to the routinely collected screening data we also present a purpose-designed, prospective cohort study of distress symptoms in the same cancer outpatient population. This study collected distress outcome scores from enrolled patients at regular intervals and with very little missing data. Consequently it contained many of the features that were lacking in the routinely collected screening data and provided a useful contrast, offering an insight into how the screening data might have been were it not for the limitations. We evaluate the extent to which it was possible to reproduce the clinical study results with the analysis of the observational screening data.

Lastly, using the modelling strategy previously developed we analyse the abundant screening data to estimate the prevalence of depression in a cancer outpatient population and the associations with demographic and clinical characteristics, thereby addressing important clinical research questions that have not been adequately studied elsewhere. The thesis concludes that analysis with observational healthcare data can potentially be advanced considerably with the use of flexible and innovative modelling techniques now made practicable with modern computing power.

EXTENDED SUMMARY

The work presented in this thesis seeks to determine if routinely collected clinical data can be used to address research questions that would ordinarily be addressed using purpose-designed clinical studies which are typically expensive and time consuming to conduct. The thesis utilises data from two sources; the first being a prospective clinical study investigating the persistence of psychological distress in cancer patients, titled the Persistence of Distress Study (the POD Study), while the second source comprises routinely collected data from a depression screening service which ran in selected oncology outpatient clinics in Scotland, UK, from 2008 to 2011. The depression screening database contained records on over 20,000 patients with repeated psychological symptoms data. Patients' clinical and demographic information from the Scottish National Cancer Registry were linked to their symptoms screening data and prepared in anonymised form by NHS Scotland Information Services Division.

Study 1: Screening patients for distress in a hospital clinic setting

Within a couple of weeks of the clinic visit the depression screening service routinely contacted patients who had scored greater than or equal to 15 on the HADS to give them a structured interview for depression. However, a concern was that the hospital environment in which the HAD scores were obtained might have added to patients' feelings of stress and could as a consequence have artificially inflated the HAD scores, requiring more patients than necessary to be followed up in the subsequent telephone depression interview. In order to examine whether the HADS scores from the clinic overestimated patients' distress levels, we followed up 218 patients who had scored high on the HADS in clinic and asked them to complete the HADS one week later over the telephone at home. We found that 72.5% (95% CI: 66.6 to 78.4%) of patients remained high scorers on the follow-up assessment, and the mean score at follow-up was 1.74 (95% CI: 1.09 to 2.39) units lower than the scores obtained in clinic. The chapter focusses on estimation of the regression to the mean effect (a statistical artefact of the sampling technique) and concludes that this accounted for most of the drop in scores between the two assessments. The study

further concludes that measuring distress on self-rated paper questionnaires in clinic is a reasonably reliable method for identifying patients with symptoms of distress.

Study II: A purpose-designed study to investigate the persistence of distress in cancer outpatients

People living with a cancer diagnosis are at increased risk of psychological distress, with some presenting as clinically depressed and requiring treatment. Little is known however about symptoms of significant psychological distress in people who do not meet the criteria for major depression. The POD Study aimed to investigate the persistence of distress over a seven month period among this population and to identify characteristics that were predictive of persistent distress. This was a prospective, purpose-designed clinical study of 325 cancer patients who had been identified by the depression screening service as presenting with symptoms of significant distress (Hospital Anxiety and Depression Scale (HADS) total score ≥ 15) at a clinic appointment. Enrolled participants were asked to complete the HADS over the telephone at regular intervals over a seven months period. Results suggested that a significant proportion of patients (37%; 95% CI: 31.4 to 42.5%) remained distressed after seven months, and that distress status one month after the initial clinic visit was a strong predictor of its persistence.

Study III: Analysis of the routinely collected screening data to address the aims of the POD Study

The large, routinely collected data from the depression screening service were analysed to determine the possibility of addressing the same research questions posed by the POD Study. We identified cases in the screening database who had scored 15 or more on the HADS at a clinic visit (the qualifying visit) and included in the analysis any available HADS data from subsequent clinic visits over the following seven months. To match the structure of data collection in the POD Study, the data were categorised into time windows. Patients' HADS scores could only be observed on their attendance at clinic appointments, and the amount of observed data was therefore linked to the frequency with which patients attended clinics. The screening

data were analysed using missing data methods coupled with flexible statistical modelling techniques using WinBUGS**.**

The estimated distress prevalence fell to just below 50% after the qualifying clinic visit and remained there with little variation thereafter. The patterns of change over time in the distress prevalence and mean distress scores were very similar over time in both the screening data and in the POD study, but the prevalence estimates were somewhat higher in the screening data. The two analyses agreed that there were no associations of persistent distress at seven months with gender, age and cancer type, and that distress status at one month after the qualifying visit was a strong predictor of persistence. There were no directly contradicting results arising from the two analyses. In the light of the limitations of the screening data the findings from the two datasets were remarkably similar.

Study IV: Analysis of the screening data to address a novel research question

In a final application of the screening data we aimed to estimate the prevalence of depression among the five most common cancer types and to identify demographic and clinical characteristics associated with depression. We followed a modelling strategy similar to that previously developed for analysis of the incomplete screening data and found prevalence rates of 13.1% (95% CI: 11.9 to 14.2%) in lung cancer patients; 10.9% (95% CI: 9.8 to 12.1%) in gynaecological cancers; 9.3% (95% CI: 8.7 to 10.0%) in breast cancer patients; 7.0% (95% CI: 6.1 to 8.0%) in gastro intestinal cancers and 5.6% (95% CI: 4.5 to 6.7%) in the almost exclusively male genitourinary cancer patients. We found that female gender, deprivation and young age were strongly associated with depression.

Conclusion

Our work demonstrates that original research with observational healthcare data is possible, with a good understanding of the clinical context, the background and the limitations of the data, when used in conjunction with flexible modelling techniques.

CONTENTS

ABBREVIATIONS

HADS	Hospital Anxiety and Depression Scale
MAR	Missing At Random
MCAR	Missing Completely At Random
MDD	Major Depressive Disorder
MI	Multiple Imputation
ML	Maximum Likelihood
MNAR	Missing Not At Random
PODS	Persistence Of Distress Study
RTM	Regression To the Mean
SCID	Structured Clinical Interview for DSM-IV
SMS	Symptom Monitoring Service

1 INTRODUCTION

Clinical practice should be based on the best available evidence. Ideally, such evidence is obtained from rigorous, purpose-designed clinical studies such as randomised controlled trials and prospective cohort studies. But gathering information in this way requires a massive effort, can be prohibitively expensive and is very time consuming. Controlled experiments are not always practicable or ethical, and sometimes answers are just needed quickly. For example, after the *Lancet* publication of the now discredited study which suggested a link between the MMR vaccine and autism (Wakefield et al., 1998) many parents opted not to have their children immunised, and answers were urgently needed to avoid putting children's health at risk. Establishing long-term links between exposure and outcomes can be notoriously difficult, and often retrospective healthcare data offer the best evidence there is. Thanks to the now widely computerised, networked data entry and storage systems in the health sector, observational healthcare data are routinely collected and stored as part of normal clinical practice. But can we rely on analysis with such data to give us meaningful answers?

1.1 Context

A recent special issue of *Statistical Methods in Medical Research* focussed on effectiveness and safety research with observational healthcare data. There are some considerable advantages of using observational healthcare data for research. Overhage & Overhage (2013) note that such data allow interventions and conditions to be studied that would not be economically feasible to evaluate in randomised controlled trials. Further, analysis with observational healthcare data often involves much, much larger samples, but at a fraction of the cost. Purpose-designed research studies, such as trials, typically use strict inclusion and exclusion criteria for patient involvement, and patients have to agree to participate. This means that the patient samples studied tend to be more homogenous. Consequently, findings may have greater generalizability when based on observational healthcare data. Such data also allow healthcare practices and treatments to be studied as they are delivered and can therefore provide better insight into real world clinical practice. In the same issue Ryan (2013) notes that observational data are interesting for safety studies because

they may provide the best evidence available for the study of rare safety events associated with medical products already on the market, and Le et al. (2013) propose a semi-automated method for rapidly evaluating safety signals using multiple longitudinal healthcare databases.

Of course, there are also many limitations to observational healthcare data. Overhage & Overhage (2013) review in broad terms the central limitations and consider the advantages and disadvantages of using administrative claims data (mostly relevant to the US) and data from clinical systems including electronic medical records. The paper lists several major challenges: Claims data may contain intentional inaccuracies introduced through efforts to maximise reimbursement. Patients may receive health care that is not logged anywhere. Absence of evidence of diagnostic tests, symptoms and treatments etc. could mean that such events did not occur, or that they were not recorded. Clinical systems are dynamic in the sense that certain events may have been routinely logged for a while, but then dropped from the logs later on. It may equally be the case that certain clinical events occurred routinely for a while before being stopped. Often there is no indication in the data itself about such procedural changes.

These authors also note that routine data are incomplete in the sense that data are typically collected when patients have an appointment, e.g. to have diagnostic tests performed, or when some other clinical procedure or check-up is needed, rather than on any regular basis. Routine data are almost invariably subject to missingness and may not therefore reliably contain the information needed for certain analyses. Second, certain scenarios or meanings may not be immediately clear from the available data so that one must rely on a combination of data to understand the meaning. For example, the authors note that: *"observational data consist primarily of transactions that occur during the care of a patient which means that there is rarely data about when a patient stopped taking medications or when a condition was resolved*."

Another important issue is that the populations included in observational data vary widely depending on the method of data collection and the setting where data were

obtained. The samples may not therefore be representative of the target population of interest. Observational healthcare data are biased in a way that controlled experiments are not since patients and their doctors choose the timing, the type and intensity of the care received. In conclusion the authors emphasise the importance of paying close attention to the detailed characteristics of the data.

Yet another article in the same issue focusses on data linkage methods for ensuring data quality. The paper by Li X & Chen (2013) reviews such methods and suggests that probabilistic matching methods using latent class models which have been applied to diagnostic testing can be translated and applied in record linkage. The paper concludes that more research is needed to evaluate the performance of various matching approaches and that there are tremendous opportunities for statisticians to collaborate with medical informaticians and computer scientists to progress research and practice in the field. Li L et al. (2013) provide a review of Inverse Probability Weighting methods and place this in the context of analysis with secondary healthcare databases for medical research, and Danaei et al. (2013) present a comprehensive analysis using observational data to investigate the effect on the risk of coronary heart disease of initiating treatment with statins. The general idea is to emulate a hypothetical randomised trial by imposing similar inclusion criteria on the analysis sample as would have been appropriate for eligibility in a trial. The emulated trial then proceeds by comparing events in patients who subsequently initiate treatment with events in those who do not. Assuming that all important confounders are known and have been measured, a valid effect estimate can be obtained by adjusting for all confounders in the final analysis. The paper concludes that the analysis yielded surprisingly promising findings and that meaningful analysis of observational data can be achieved with the right combination of high quality data, good subject-matter knowledge, and appropriate statistical methodology.

1.2 Motivation for the present study

Analysis with routinely collected, observational healthcare data is clearly an important and topical area of research. We had access to repeated psychological symptom data that were routinely collected from over 20,000 outpatients who

attended selected oncology clinics in Scotland, UK. Patients attending for appointments were approached in the clinic waiting area by a symptom monitoring and screening service managed by our group, Psychological Medicine Research at the University of Edinburgh. The screening service asked patients to complete a questionnaire which enquired about physical and psychological symptoms. The questionnaire answers were used to help the oncologists address issues that were of concern to the patients, and also to identify patients with major depression who were eligible for enrolment in the SMaRT oncology-2 and 3 trials (Walker, Cassidy & Sharpe, 2009a, 2009b). However we wondered if these routinely collected data might also offer a unique opportunity to study the progress of distress symptoms on an unprecedented scale in this population. Clearly the limitations were many, so how could we make the most of these data?

Separate from the depression trials Psychological Medicine Research also conducted a purpose-designed, prospective cohort study into the persistence of distress symptoms in the same cancer outpatient population. The Persistence Of Distress Study (PODS) collected distress outcome scores from enrolled patients at regular intervals with very little missing data and consequently contained many of the features that were lacking in the routinely collected screening data. Perhaps PODS could act as a useful contrast to the screening data, a sort of gold-standard offering an insight into what the screening dataset might have looked like were it not for its limitations. Would we be able to reproduce the results obtained with PODS through secondary analysis with the observational screening data? Clinical cohort studies have limitations of their own. Could there even be advantages to the abundant but highly irregular screening data that might actually outweigh those of the purpose-designed, but much smaller clinical study?

These were the specific circumstances which motivated the project. Anchored in these concrete data and questions the work described in this thesis aims to investigate how meaningful research using observational healthcare data might proceed.

1.3 Scope

Chapter 2 outlines some early and exploratory work before placing the problem in a missing data context. Relevant methods for handling missing data are further reviewed in Chapter 3. The focus is returned to the routinely collected symptom data in Chapter 4. The chapter considers the intra-patient correlation between repeated distress measurements and examines the extent to which patient selection criteria employed in the POD Study (and routinely by the depression screening service to identify patients that are likely to be depressed) induce regression to the mean in subsequent outcomes. The POD Study and its findings are the subject of Chapter 5. Bayesian modelling with incomplete data is introduced in Chapter 6. The chapter provides an account of exploratory work with different modelling approaches in WinBUGS to analyse incomplete data that are simulated under known missingness processes using the SAS software. In Chapter 7 we construct a dataset from the routinely collected symptom data that mimics the POD Study data structure. We then develop a Bayesian model in WinBUGS to estimate the posterior predictive distribution for the incomplete longitudinal response data and covariates under both Missing At Random (MAR) and Missing Not At Random (MNAR) mechanisms and use this model to generate multiply imputed datasets for further analysis. The findings are compared with those from the prospective POD Study. Finally in Chapter 8 we follow a similar modelling strategy using the observational screening data to address clinical research questions that have not been adequately studied elsewhere. The present research project is summarised in Chapter 9. We reflect on the objectives, the wider context, limitations, strengths and future direction.

2 BACKGROUND

This chapter is introduced with an overview of an early review of the literature on analysis with routinely collected patient health outcomes. This was not the result of a formal systematic review and should not be seen as providing an exhaustive account of the available literature on the subject. Rather it serves to define the topic by providing an impression of the breadth of this, and to put into context the current project. The project is then placed within a missing data context, and lastly we describe Rubin's taxonomy for classifying mechanisms for missing data.

2.1 Review of the background

Routinely collected patient health outcomes include physiological measurements, psychological symptoms and self-reported functional scores, quality of life measurements, patients' satisfaction with care etc. Collecting such patient outcomes has numerous purposes that can broadly be divided into two categories. Firstly, at the patient level, individual responses may be used for the screening or monitoring of outcomes in individual patients. The routine collection of general health outcomes from patients also facilitates communication about the patient within medical teams, as well as directly with the patient, thereby promoting patient-centred care (Greenhalgh, 2009). Secondly, at the group level, the aggregate data from all patients may be used as a means of measuring the quality of a clinical service, assessing the effect of an intervention on patient outcomes or to research other epidemiological questions.

Rose & Bezjak (2009) consider the logistics of collecting patient reported outcomes in a clinical setting and draw attention to a number of factors that are necessary for the successful collection of useful data. The authors emphasise the importance of having routine procedures in place to minimise data that are incomplete or biased towards more compliant patients. Analyses based on routinely collected outcomes should be interpreted with caution. Davies & Crombie (1997) provide a detailed account of the pitfalls in interpreting such outcomes. Contrary to controlled experiments where patients are allocated at random to interventions and test procedures, we can exercise no such control with routinely collected patient

outcomes. Patients are not sampled with a specific research question in mind, and the sampled patients may not be representative of the research target population.

Systematic non-response in routine medical outcomes can produce data that are biased for example towards more treatment compliant patients, or patients with particular demographic or health characteristics. The limitations are exacerbated when considering routine data for longitudinal research. For convenience, patient outcomes are sometimes collected immediately before a patient's clinical appointment. As a result, patients' outcome measurements are inseparably linked with the event of having a clinic visit. The frequency of clinic visits (and therefore the amount of observed data) can vary between patients depending on the health state and, if in treatment, the treatment stage of the patient. A cross-sectional analysis of outcome data might focus on data from the first visit from each patient, thereby ensuring that all sampled patients contribute equally to the analysis. In a longitudinal analysis, the frequency with which data are observed (and therefore the amount of data observed from each patient) may have a direct bearing on the value of the outcomes. Routinely collected longitudinal healthcare data might be valuable but simple analysis strategies are unlikely to suffice.

The analysis of routinely collected outcome data in healthcare is a broad topic. Analysis of outcomes have been used (and misused) by policy makers and healthcare managers as a tool for monitoring targets and managing the performance of clinical staff (Epstein, 1990; Blisker & Goldner, 2002; Lilford et al., 2004). With the increase in electronic databases used for storing patients' medical records there has been a growing interest in the reuse of data extracted directly from medical notes for clinical research (Curcin et al., 2010; Rosenbloom et al., 2012). A recent publication in the *Lancet* studying child maltreatment across six different countries used longitudinal analysis of routine healthcare data to investigate the scale of the problem and trends over time (Gilbert et al., 2012). However despite growing interest in analysis with such data relatively little methodological work has been published on the topic. There are a few published studies on the use and analysis methods of routinely collected clinical data within the mental health literature (e.g. Leaf et al., 1993,

Macdonald, 2002). Interrupted Time-Series analysis of routine longitudinal healthcare data has been proposed as a pragmatic alternative to expensive and time-consuming randomised controlled trials in places and situations where such trials are not feasible, for example to evaluate changes in health policy in developing countries (Lagarde, 2012). However, time-series analysis is best suited to the modelling of many correlated data points from a single (or very few) series such as is commonly observed in economics. These data structures are less common in clinical studies which are typically concerned with many independent units (usually patients), each contributing with only a few correlated measurements. We found a number of other studies on broadly related topics. Sithole & Jones (2003) use GP prescription data to fit a Bayesian repeated measures model for detecting differences in GP prescribing habits following an educational intervention. Their data are complete at every time point and the paper does not consider some central limitations to routinely observed data. Hogan & Lancaster (2004) use the example of an observational HIV natural history study to compare instrumental variables (IV) methods, typically used in social sciences and econometrics for drawing causal inferences in the presence of potential unobserved confounders, with the epidemiologic approach of inverse probability weighting. McDonald et al. (2009) report on a population-based record-linkage study of mortality rates in people with Hepatitis C in Scotland based on national databases. The focus of this study is predominantly on data linkage rather than statistical modelling. Finally, He et al. (2010) use multiple imputation to complete a centralised cancer care outcomes survey dataset with considerable missingness on multiple variables for subsequent multi-objective analysis by external investigators.

2.2 Choosing an angle

There are many different ways of handling irregular spacing in the analysis of longitudinal data. We considered some alternative approaches in exploratory work. Lowess curves and various other plots were produced to summarise the distress scores from the observational symptom screening data in continuous time. The dependency between repeated measures could be induced implicitly by including random effects into the linear model. We experimented with linear spline models

with random intercepts and slopes. However we required a modelling framework that would be suitable also for the POD Study data and its research aims. One solution was to think of the unevenly spaced observations as an incomplete picture of a hypothetical, complete dataset consisting of regular and frequent values. The set of complete data from each patient then consisted of both observed and unobserved values that in principle should have been observed. Replacing the idea of irregularly spaced data with a mental picture of a partially observed, complete dataset allowed us to approach the problem with methodology developed for analysis with missing data.

2.3 The missing data framework

Analysis in the presence of missing data broadly presents three types of challenges (Fitzmaurice, Laird & Ware, 2002). Firstly, with missing data there is a loss of information which will result in less precise estimates (i.e. larger standard errors, wider confidence intervals and larger p-values) and analyses are therefore more likely to be inconclusive. Secondly as the number (and timing) of observations will vary between patients, missing data impose restrictions on the methods of analysis that may be applied as not all methods will handle unbalanced data. Finally, and perhaps most importantly, missing data may compromise the validity of the data and careful thought should be given to the mechanisms that cause the data to be missing.

2.3.1 The missing data mechanism

The type of assumptions that can be made about the missing data has important implications for the appropriate analysis strategy necessary for valid inference. Rubin (1976) developed the taxonomy for the missing data mechanism. This general framework can be extended to the case of repeated data (e.g. Laird, 1988). Most work on methods for analysis with missing data since the introduction of Rubin's taxonomy has used the general idea behind this taxonomy to classify the mechanisms for missing data. The following description of three types of missingness mechanisms relevant to incomplete longitudinal data is loosely based on Fitzmaurice, Laird & Ware (2002). These authors distinguish between Missingness that is

Completely At Random (MCAR), Missingness At Random (MAR) and Missingness that is Not At Random (MNAR) as described below.

Missing completely at random

Assuming all patients have at least one observed value, a missing value can be thought of as missing completely at random (MCAR) when the probability of missingness is unrelated to both observed responses from the same patient as well as the values of unobserved responses that in principle should have been observed.

If Y_i denotes the complete vector of all responses for the ith subject we can partition this into two vectors $Y_i = (Y_i^O, Y_i^M)$, where Y_i^O is the vector of the observed responses and Y_i^M is the vector of unobserved responses that in theory should have been collected. Associated with Y_i is a vector R_i of missing data indicators (R_{i1}, R_{i2}, …) where R_{ij} is equal to 1 when the *j*th response for subject *i* is observed and zero otherwise.

The MCAR assumption is then satisfied when the distribution of R_i (conditional on the covariate data X_i) is independent of all observed and unobserved responses.

$$f(R_i \mid Y_i^O, Y_i^M, X_i) = f(R_i \mid X_i)$$

When data are MCAR the governing distributions of the observed, Y_i^O, and the complete data, Y_i, are the same. The means, variances and covariances of the observed data are equal to those of the complete data on average.

The important implication of this is that inferences can be drawn about the distribution of the complete data from moments of the observed data alone. As a result it is possible to analyse the observed data as if it were the complete data.

Missing at random

If the probability of a value being missing is unrelated to values of the unobserved responses, but not necessarily independent of other *observed* responses from the same patient, the data are said to be missing at random (MAR). Thus, missingness completely at random is a special case of missingness at random.

Data are MAR when the conditional distribution of R_i (conditional on covariate data) is independent of the unobserved responses, but not necessarily of the observed responses.

$$f(R_i \mid Y_i^O, Y_i^M, X_i) = f(R_i \mid Y_i^O, X_i)$$

When data are MAR the distributions of the missing and observed responses are identical within strata of the data defined by covariates and the vector of observed responses. By analysing the observed data appropriately it is therefore possible to draw inferences about the complete distribution of Y_i from the observed data alone.

When data are no longer MCAR more care will have to be exercised in the analysis of the data. In particular, the moments of the observed data are no longer equal to those of the complete data. Consequently, biased estimates will result from a naïve analysis of only the complete cases or in any analysis of the available data (see below) that does not adequately account for the missing data.

Missing not at random

When data are Missing Not At Random (MNAR) the probability of missingness is dependent on the values of the unobserved responses. Under MAR it is possible to model the complete data without having to specify the distribution for the missing data indicators R_i. This is not the case when the probability of an observation being missing is related to the value of the unobserved responses. That is, even after stratifying by all the observed values (and covariate data), missingness within a particular stratum of observed values is no longer random but is dependent on the unobserved data itself. In this case it is necessary to specify the form of the

dependency between the missing data indicators R_i and the unobserved responses Y_i^M. Careful attention to the cause of missing data is required especially in the presence of a large amount of missing data that are MNAR. The probability of an observation being missing can be modelled, for example, in a logistic regression model as a function of both the observed and unobserved responses (Diggle & Kenward, 1994). The joint likelihood is then a function of both the outcome and missingness processes. The nature of the relationship between the missing data indicators and the unobserved responses cannot be verified from the data and it is not generally possible to know with certainty the parameters governing the missing data mechanism. Analysis under a MNAR mechanism therefore often focuses on sensitivity of conclusions reached in a main analysis to alternative scenarios for the missingness process. Little (1995) discusses two commonly used methods for modelling the data and the drop-out mechanism together, selection models and pattern mixture models.

3 METHODOLOGY

The previous chapter introduced a formal framework for characterising missingness in longitudinal studies. We will now provide details of some of the prevailing methods for analysis with incomplete multivariate data. The chapter is introduced with a brief overview of some common approaches to handling missing data. The main focus of the chapter will be on more recent, principled approaches to modelling incomplete data under named assumptions for the missingness mechanism.

3.1 Common approaches to analysis with missing data

Traditionally many analysis methods required a rectangular dataset to proceed. This meant that incompletely observed cases were a nuisance to analysts. Many conventional methods for handling missing data, sometimes referred to as ad hoc methods, were therefore primarily concerned with generating analysable datasets and less concerned with the underlying missingness mechanism and potential bias present in the incomplete sample.

Complete-case

Perhaps the simplest way of dealing with the nuisance of incomplete cases in longitudinal studies is to omit all patients with any missing data from the analysis. The resultant dataset can then be analysed using conventional complete-data methods. Although appealing due to its simplicity the use of this method, also known as list-wise deletion, is generally not regarded as good practice and is commonly criticised (e.g. Fitzmaurice, 2003; Schafer & Graham, 2002; Little & Rubin, 2002, Chapter 3). Except from situations where missingness is MCAR, the distribution of data from patients with complete data will not in general be identical to that of the study population and the method can yield potentially seriously biased results. Analyses based on this method are also typically very inefficient, especially in studies where there are many follow-up occasions or several outcome variables that are subject to missingness.

Available-case

A more efficient use of the data is through methods that will use all available data. Such methods, referred to as *available-case* methods, are capable of handling unbalanced datasets, so that patients with partially observed responses still contribute with data where it is available. The method of Generalised Estimating Equations (Liang & Zeger, 1986) can be thought of as an available-case method. The assumption about the missingness mechanism underlying this class of methods is still that of MCAR. That is, analysis with missing data through available-case methods will generally only provide unbiased results when missing data are not related to other observed values from the same subject, nor the value of the missing responses themselves.

Imputation methods

A commonly used alternative to complete-case and available-case methods is to use one of the many methods of imputation to generate a complete dataset thereby avoiding wasting information belonging to patients where some measurements are unobserved. The basic idea behind these techniques is to replace missing observations with plausible values, and then use standard methods of analysis on the complete (imputed) dataset.

The simplest imputation methods replace a missing value from a patient with a single imputation, for example using the mean of a patient's observed values. In clinical studies when some values are unobserved due to drop-out another common imputation method is the Last Observation Carried Forward method (LOCF) whereby unobserved measurements are simply filled in using a patient's last observed value. The implicit premise is that a patient's last observed value provides a good estimate of the subsequent unobserved outcomes over the remaining follow-up time. Although widely criticised (e.g. Ware, 2003; Carpenter et al., 2004; Mallinckrodt et al., 2004) this method is still used in clinical trials of pharmaceutical products. (It should be noted that methods such as LOCF are not inherently wrong. For example, there might be situations where the researcher genuinely believes that LOCF describes a plausible mechanism for the missing data.) Although known not in

general to be the case, it is sometimes argued that LOCF is conservative. It is argued that the method assumes that the effect of the trial intervention is less in patients who drop-out. In fact it is not hard to think of situations where the use of LOCF will bias estimates in favour of a trial intervention. Molenberghs & Kenward (2007, pp. 47-49) compare the complete case and LOCF methods in the context of missingness mechanisms. They show that, in general, effect estimates from either analysis are biased under a MAR mechanism and that LOCF generally produces biased estimates even under a MCAR mechanism with the bias working in either direction depending on the parameters involved. The authors conclude that it is not immediately clear what conditions need satisfied for the LOCF method to be unbiased or conservative.

Of course there are other strategies for imputation that can be employed. It might be assumed that the (unmeasured) value of the outcome variable is zero when a patient's measurement is unobserved or that the outcome variable returns to baseline levels when a patient drops out. Unconditional mean imputation uses the overall average of observed scores from other patients at the follow-up time being imputed. Several other variants of mean imputation exist where the aim is to use the existing data in an optimal way to best predict the missing values. Fitzmaurice (2003) provides a brief overview of more advanced techniques that use regression methods to fill-in a patient's missing values using the conditional means (conditional on observed responses). The incomplete variable is regressed on all other observed variables and the resultant regression coefficients, together with the patient's observed responses are then used to predict the missing values. Usually a random term is added on from a simulated normal distribution with variance equal to the residual error variance in the regression model. This is to avoid the problem of small variances in the imputed values mentioned above. Other methods include predictive mean matching where the missing value is imputed using an empirical value from the observed data that is close to the predicted value (e.g. Little & Rubin, 2002) and propensity scores methods (Rosenbaum & Rubin, 1983) that associate a probability (or propensity) of drop-out with each patient and impute missing observations using values from patients with similar propensity scores.

However even if the analyst succeeds in producing imputations that on average are unbiased under a MAR mechanism, a conventional analysis based on singly imputed data will still produce estimates with overstated precision (Schafer & Graham, 2002). That is, the standard errors will be too small and confidence intervals too narrow. The random element inherent in producing the imputed values is unaccounted for since the imputed observations are analysed as if they had actually been observed.

This problem which is common to all single imputation prompted Rubin's work on Multiple Imputation (MI) in sample surveys (Rubin, 1977, 1978, 1987) marking the beginning of extensive work on multiple imputation methods. Multiple imputation has been researched and developed extensively since Rubin's initial work, and still occupies a very important role in the current literature on principled approaches to analysis with missing data. The theory of MI is also integral to much of the analysis in the present project and will be covered in detail below.

Likelihood based methods

When data are MAR valid inference can also be gained from maximum likelihood (ML) estimation based on the full joint distribution of the multivariate responses. In the presence of missing data, this method also produces estimates that account for the loss of information (or uncertainty arising from the missing observations) and will produce correct standard errors. Likelihood-based methods require correct assumptions about the joint distribution of the repeated data. These methods will not in general provide valid estimates in the presence of missing data if the likelihood function is incorrectly specified, for example if the covariance structure is not modelled correctly. When there is missing data, and the missing data pattern is non-monotone (section 3.4.6), maximising the likelihood function algebraically is not usually possible. Generally in this situation, no closed-form expression exists for the maximum likelihood estimate and maximisation will have to be done through an iterative and computationally intensive procedure. Expectation-Maximisation (EM) algorithms (Dempster, Laird & Rubin, 1977) can provide maximum likelihood parameter estimates from incomplete data by alternately estimating the unknown parameters, and the missing values, in an iterative algorithm.

Little & Rubin (2002) provide a comprehensive reference text on missing data methods and divide these methods into four broad classes: Complete data methods such as the list-wise deletion methods; weighting methods that use algorithms to compensate for observations that are essentially underrepresented in the observed sample; imputation-based methods; and finally model-based methods. We will be concerned with principled approaches such as imputation-based (particularly multiple imputation) and model-based methods (also known as direct likelihood or joint modelling methods). For now however we shall return briefly to the building blocks of the missing data paradigm.

3.2 Ignorability

We saw in Chapter 2 that incomplete longitudinal data may be classified according to Rubin's taxonomy as MCAR, MAR or MNAR. An important related concept is that of ignorability (Rubin, 1976; Little & Rubin, 2002, Chapter 6). Informally, when the data are MCAR or MAR and we wish to make inference about the parameters governing the response process, then the missingness function may be ignored under likelihood-based inference. (Strictly, a further requirement for the ignorability condition to be satisfied is that the parameters governing the response and the missingness processes are distinct.)

In general we can define the complete set of data arising from a study that is subject to missingness as consisting of Y^O, the observed responses, Y^M, the unobserved responses and R, the missing data indicators. The missing data indicators are simply binary variables that record whether a given response, y_{ij} say, was observed ($r_{ij}=1$) or missed ($r_{ij}=0$).

Having defined the complete set of data in this way we can now define the complete-data likelihood as a joint function of the responses $Y=(Y^O, Y^M)$ and the missing data indicators R

$$L_{full} = f(Y^O, Y^M, R \mid \theta, \psi)$$

Here θ are the parameters governing the response process and ψ are the parameters governing the missingness process.

The likelihood can be decomposed in terms of conditional probabilities

$$f(Y^O, Y^M, R \mid \theta, \psi) = f(R \mid Y^O, Y^M, \psi) f(Y^O, Y^M \mid \theta)$$

The first part of this likelihood is the function governing the missingness process (conditional on the responses and ψ). The second part is the function defining the response process.

In most cases we are concerned with estimating θ whereas ψ and the functional form of the missingness mechanism typically are not of direct interest. The ignorability condition implies that the full-data likelihood (with respect to θ) is proportional to the partial likelihood consisting of the response process on its own (i.e. ignoring the missing data process). In other words, when the missing data process is ignorable there is no additional information about θ contained within $f(R \mid Y^O, Y^M, \psi)$, the part of the likelihood arising from the missingness process.

We saw in Chapter 2 that MAR missingness implies that

$$f(R \mid Y^O, Y^M, \psi) = f(R \mid Y^O, \psi)$$

by definition. When the missingness process does not depend on the Y^M then there is no information about θ to be extracted from the missing data indicators. The likelihood arising from the missingness process is effectively constant over the parameter space for θ. Direct likelihood estimates for θ under the ignorability assumption can then be obtained in the usual way by maximising the marginal likelihood for the Y^O having integrated out the Y^M:

$$f(Y^O \mid \theta) = \int f(Y^O, Y^M \mid \theta)\, d\, Y^M$$

3.3 The direct likelihood method

Sometimes referred to as the model-based or joint modelling approach, the method of direct likelihood is one of the central principled methods for handling incomplete longitudinal data. When the data are MAR, and the likelihood function for the non-response mechanism is ignorable, one can obtain unbiased inferences about the parameters governing the response mechanism, θ, "simply" by maximising the observed-data likelihood, $f(Y^O \mid \theta)$. Generally when the data are not MAR, one is required to maximise the full-data likelihood $f(Y^O, Y^M, R \mid \theta, \psi)$ to obtain valid inference about θ. We return to this latter scenario in section 3.6. For now we will focus on the situation when data are MAR.

Generalised Least Squares (GLS) regression which is the multivariate generalisation of Ordinary Least Squares Regression will produce estimates for the means and precisions that are biased when data are not MCAR. This is not generally the case when using distribution based regression methods. Molenberghs & Kenward (2007, pp. 50-52) illustrate how likelihood-based estimation of mean parameters under a bivariate normal distribution are based on the conditional expectation of unobserved responses given the observed responses. GLS estimates however do not draw on distributional assumptions to utilise information from the whole sample, but rely simply on the isolated information available for each parameter.

In practice of course, maximising $f(Y^O \mid \theta)$, the observed-data likelihood, is not generally straightforward. When the missing data pattern is monotone it is possible to construct closed-form expressions for the maximum likelihood estimate (Little & Rubin, 2002), however with non-monotone patterns we are generally required to use numerical maximisation techniques. The Expectation-Maximisation (EM) algorithm (Dempster, Laird & Rubin, 1977) is widely used for maximising likelihoods arising from incomplete data and is commonly implemented in software procedures capable of likelihood-based longitudinal and multivariate analysis.

Consequently, the use of the direct likelihood method for handling missing data typically requires very little effort on the part of the analyst since most modern statistical packages use procedures such as the MIXED procedure in SAS that easily accommodate incomplete data records and use numerical methods for maximising the likelihood.

When using the direct likelihood method to obtain parameter estimates from incomplete data, it is important to assess the modelling assumptions. In principle it is the case that parameter estimates are only valid when the multivariate distributional assumptions are correct. To what extent parameter estimates are impacted when the assumptions do not hold depends on the amount of missingness and the degree of deviation from the distributional assumptions. When data are MAR the maximum likelihood estimates of the mean parameters can be expressed as a function of the variance-covariance parameters; it follows that misspecification of the covariance matrix structure will impact on the mean parameter estimates. Finally, misspecification of the linear predictor, whether through the functional form, omission of endogenous variables or variables related to both missingness and the response, can equally result in biased estimates.

We have seen that it is possible to make valid inferences about the response mechanism from incomplete MAR data while ignoring the missingness mechanism. It should be noted that it may still be desirable to model the missingness mechanism in these circumstances if understanding this process adds to the overall value of the study findings.

3.4 Multiple Imputation

Some common strategies for simple imputation were discussed above. Many of these posited assumptions about the missing data that may not be realistic in many situations. But even where single imputation methods succeed in predicting missing values correctly the problem still remains that conventional analysis of singly imputed data tends to overstate precision in the estimates by not taking account of the inherent uncertainty in the imputations. There are different ways of handling this

problem in singly imputed data (Little & Rubin 2002, Chapter 5) although Multiple Imputation (MI) has become the method of choice. Below we give an account of MI methods under an ignorable missing data mechanism.

Under MI missing values are imputed *multiple* times through an imputation model. The imputation model is implemented within a Bayesian imputation scheme which generates predictions of the unobserved data that are valid under some named assumption for the missingness mechanism. Unlike single imputation where analysis is based on a single imputed dataset, MI produces a number of completed datasets to be analysed in combination. The datasets each consist of an observed part and an imputed part where the observed parts are identical across datasets but the imputed parts vary. Once the imputation step is complete each imputed dataset is analysed separately using the appropriate methods that would have been applied had the data been fully observed. Finally the results of the separate analyses are combined using the rules formulated by Rubin (1987) to produce overall (pooled) estimates with correct standard errors that take account of the variation between the imputed datasets.

3.4.1 The analysis step

3.4.1.1 Rubin's rules: combining the imputations

For the single-parameter case Rubin (1987) developed the following general rules for pooling the point estimates (and associated errors) across *m* imputed datasets to form a single estimate with correct standard error that accounts for both within and between imputation error.

Suppose some parameter A is estimated separately with each of the m imputed datasets and denote these estimates $\tilde{A}_1, \tilde{A}_2, ..., \tilde{A}_m$. An unbiased pooled estimate of A is simply

$$\hat{A} = \frac{1}{m}\sum_{i=1}^{m} \tilde{A}_i$$

the mean of the m individual estimates. Now, let V_i denote the variance of the parameter estimate from the ith imputed dataset. Rubin's variance formula then gives the pooled within-imputation variance as

$$V = \frac{1}{m}\sum_{i=1}^{m} V_i$$

The between-imputation variance is simply estimated as

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\tilde{A}_i - \hat{A}\right)^2$$

And the total variance of $\hat{A}$ is then

$$T = V + (1+m^{-1})B$$

The $(1+m^{-1})$ factor corrects for the uncertainty arising from having a finite number of imputations. Therefore as m approaches infinity the total variance simply reduces to the sum of V and B, the within- and between-imputation error, in which case the use of the data is fully efficient in the sense that all the information available in the incomplete data about the parameter is utilised. Hypothesis tests are conducted by comparing the usual test statistic

$$\frac{\hat{A} - A}{\sqrt{T}}$$

to the appropriate t distribution. The degrees of freedom here depend on the number m of imputed datasets as well as the sample size. However with large m this is less of an issue and comparisons can be made with the conventional reference distribution instead.

These results readily generalise to the multi-parameter case (Little & Rubin, 2002, p. 211). A is then a vector of parameters and $\hat{A}$ are the associated parameter estimates. V is the within-imputation covariance matrix and

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\tilde{A}_i - \hat{A}\right)\left(\tilde{A}_i - \hat{A}\right)'$$

is the between-imputation covariance matrix. For the multi-parameter case with large m the test statistic may be compared to the usual Chi square distribution with degrees of freedom equal to the length of the parameter vector A.

3.4.1.2 Fraction of missing information

Considering again the variance formula $T = V + (1+m^{-1})B$ we note that V, the average variance within each (imputed) dataset, is an estimate of the full-data variance. In other words, V estimates the variance of the parameter estimate in the (hypothetical) fully observed data. It is also clear that the additional variance caused by the missing data is $(1+m^{-1})B$ which reduces to B as m approaches infinity.

The relative increase in variance caused by the missing data is therefore B/V. In a sense this represents the penalty arising from having unobserved data: the variance of the parameter estimate will always be inflated by at least this amount regardless of the number of imputations generated.

Another important quantity is the fraction of missing information, λ, about the parameter to be estimated. Defining the variance components above in terms of Fisher information this can be shown to be $\lambda=B/(V+B)$. Some authors suggest using this fraction as a guide for deciding on the number of imputations needed to obtain good efficiency (relative to estimates based on infinite m) and stability in the results.

The relative efficiency of an estimate based on m imputations relative to an estimate based on an infinite number of imputations

$$\left[\frac{V+(1+m^{-1})B}{V+B}\right]^{-1}$$

can also be expressed in terms of the fraction of missing information:

$$\left[1+\frac{\lambda}{m}\right]^{-1}$$

Using this formulation it is immediately obvious that the number of imputations needed to achieve a given relative efficiency increases with the fraction of missing information about the parameter. The additional error incurred from basing estimates on a finite number of imputations is referred to as the Monte Carlo error. The variance formula can be written as

$$T = V + B + (B/m)$$

(B/m) is the variance arising from the fact that MI is a stochastic procedure that, with finite m, yields different results when repeated. The square root of this variance term $(B/m)^{1/2}$ is the Monte Carlo error.

3.4.1.3 How many imputations are needed

We saw above that the number *m* of imputed datasets affects the precision with which we can estimate an unknown parameter. *m* should be chosen such that the relative efficiency is acceptably close to unity or it may be the case that independent repeats of the analysis leads to different conclusions because of the sampling variability in separate MI runs. Historically many authors (e.g. Shafer, 1997; Rubin, 1987) have suggested using only a small number of imputations arguing that the efficiency gain diminishes rapidly as *m* increases beyond 5 to 10 imputations. However more recently it has been recognised that the reproducibility requirement necessitates a much larger number of imputations. White, Royston & Wood (2011) suggest as a rule of thumb that the number of imputations should be no less than the percentage of incomplete cases. E.g. with 26% incomplete cases we would choose

m=30. With recent gains in computational speed the number of imputations is (usually) no longer restricted by computational resources. There is however an argument for parsimony since results based on 3 to 5 datasets are arguable more transparent and easier to verify than results based on hundreds of imputed datasets.

3.4.2 The imputation model

The imputation model is the crucial link that binds together the predicted values and the observed data. As an example, suppose we had planned to observe a quantitative variable at baseline, Y_1, and again at follow-up, Y_2, some time afterwards. As is not unusual in longitudinal studies we are concerned here with the value of Y_2, the outcome at the last planned appointment. Suppose further that the baseline measurements were fully observed but that some patients failed to provide outcome data on the follow-up occasion. When the distribution of Y_2 is approximately normal the missing outcomes may be modelled through a linear model of the form

$$E[Y_2] = \alpha + \beta Y_1 + \gamma_1 X_1 + \gamma_2 X_2 + \ldots + \gamma_p X_p \qquad (3.1)$$

Here $X_1, X_2, \ldots, X_p$ are covariate data related to the outcome through the regression coefficients $\gamma_1, \gamma_2, \ldots, \gamma_p$, α is the intercept and β relates the baseline outcome to the outcome at follow-up. Broadly, the imputation procedure then consists of the following two steps: In the first step Y_2 is regressed on Y_1 and covariates; in the second step the coefficients estimated in this regression are used to predict the missing Y_2 after adding random components. Parts of the random components contribute to variability in the predictions within each imputed dataset, and parts contribute to variability between the m sets of imputed data.

3.4.2.1 Building the right imputation model

If the imputation model is misspecified in some way, for example by omitting important predictors of the missing responses, then the predictions of the missing values may be biased which in turn can lead to serious bias in the overall analysis depending on the amount of missing data and the nature of the misspecification.

Sterne et al. (2009), Spratt et al. (2010) and White, Royston & Wood (2011) set out strategies for the implementation and reporting of MI models and consider a number of scenarios including clinical trials, missing covariate data and what to do when the substantive model includes interaction and higher order terms. Generally the following variables should be considered for inclusion in the imputation model: all previously (and subsequent) observed outcomes, all variables related to the outcome being imputed and all variables predictive of missingness. The imputation model should retain all structures that are modelled in the substantive analysis such as interactions and higher order terms. In the case of imputation of missing covariate data it is important to include the outcome variable as a predictor of the covariate. This is especially true when the substantive analysis investigates associations between the outcome and incomplete covariate data since omitting the outcome from the imputation model would force the regression coefficient to be zero in the imputed data.

When the imputation and the substantive models are perfectly consistent in the sense that they include the same explanatory variables in their linear predictors and posit the same functional forms to describe the relationships between the explanatory variables and the outcomes then the two models are said to be congenial. Conversely when the imputation and the substantive models are not consistent in the manner just described the models are said to be uncongenial which is a situation of special interest (Meng, 1994). The possibility of having uncongenial models is often put forward as one of the major strengths of MI. The direct likelihood method for handling missing data requires the full specification of the analysis model, but is only useful with that one analysis. With MI the imputer can impute the data using a model that is consistent with the underlying missingness mechanism while the subsequent analyst may subject the data to a number of different analyses and summaries that were not necessarily anticipated by the imputer. Molenberghs & Kenward (2007) point to a number of situations where uncongenial models are very useful including when using multiply imputed data with Generalised Estimating Equations that are generally only valid under MCAR to obtain valid inference under MAR missingness. Uncongenial models also provide a useful tool for conducting sensitivity analyses.

This can be done for example through altering the imputation model to accommodate a non-random missingness process thereby assessing the robustness of the results from the (uncongenial) substantive model to MNAR missingness. Kenward & Carpenter (2007) refer to examples of a variety of such sensitivity analyses.

3.4.3 Making proper imputations

The term *proper imputations* refers to the idea that the imputations should not only reflect the variance present in the observed part of the data but also the uncertainty about the true underlying parameters governing the responses. Proper imputations are therefore generated by first formulating (and drawing from) a distribution for the parameters and then using each realisation of the parameters to generate one set of predictions for the missing values according to the relations defined in the imputation model. At its heart MI is a Bayesian procedure with imputations generated from the Bayesian posterior predictive distribution

$$f(Y^M \mid Y^O) = \int f(Y^M \mid Y^O, \theta) f(\theta \mid Y^O)\, d\theta$$

where $f(\theta \mid Y^O)$ is the Bayesian posterior distribution for the parameters given the observed-data likelihood. The imputations are generated by simulating draws from the predictive distribution above. In practice this is done by making draws from

$$Y^M \sim f(Y^M \mid Y^O, \theta) \tag{3.2}$$

at m different realisations of

$$\theta \sim f(\theta \mid Y^O) \tag{3.3}$$

Essentially the imputation task proceeds as follows. A random draw is made first from the posterior distribution of the parameters (3.3). Second, this realisation of the parameters is used in (3.2) to create one imputed dataset by drawing from the predictive distribution of Y^M. The two steps are repeated m times to create m imputed datasets.

As was mentioned earlier this two-step procedure for making imputations is often referred to as *proper* because it properly accounts for uncertainty in the underlying parameters. Conversely, imputations based on (3.2) with θ replaced by a fixed observed-data estimate are sometimes referred to as improper. When making proper imputations a Bayesian prior distribution has to be chosen for the parameters. Typically this is chosen to be non-informative (e.g. the SAS software uses the non-informative Jeffreys prior as a default (SAS OnlineDoc®, MI procedure)). This seems to be a sensible strategy when MI is used within an otherwise frequentist analysis.

Sometimes it is straightforward to define the Bayesian posterior distribution for the parameters in (3.3). But depending on the missingness pattern the observed-data likelihood may be a complicated function and a numerical iterative procedure is required to obtain the distribution. This situation is considered in more detail below.

3.4.4 When the response is not normally distributed

Imputations based on the imputation model in (3.1) are typically imputed under a normal distribution. However, it is common in clinical research to have outcomes that are skewed or that are measured on discrete scales, ordered categorical or binary scales. One approach to imputation in non-normal variables is simply to proceed regardless and apply a linear regression model such as that in (3.1). Many authors have reported good results using normal linear regression for imputation with highly non-normal data (e.g. Schafer 1997, Chapter 5). Misspecification of the distribution governing the imputation model affects only the imputed data and the effect of such misspecification is therefore negligible when the fraction of missing data is small. Nonetheless, there are clear limitations to this approach. Many quantitative variables take values only within a certain range. Consequently, imputations above or below the permissible range may have to be handled in some way, for example by assigning the lowest permissible value within the range to all imputations below the lower limit and the highest permissible value to all imputations above the upper limit. Alternatively, a normalising transformation could be applied to non-normal

quantitative data before using the normal linear model for imputation. Another way to ensure that the imputed data mirror the observed data is through the use of predictive mean matching methods. These ensure that the missing observations are imputed using only values that already exist within the observed part of the data. This is done by matching each individual with missing data to a number of observed individuals with similar linear predictors. The imputation is then drawn at random from one of the matches (Little, 1988).

However MI is not confined to imputations from the normal distribution. In the case of imputation with a single incomplete variable it is straightforward to use a non-normal imputation model that is compatible with the type of variable being imputed. For example, a logistic regression model may be used to impute univariate binary data and an ordinal logistic regression to impute univariate ordered categorical data.

3.4.5 Sequential imputations

The problem of modelling different types of variables as described above is more challenging in the *multivariate* setting. When more than one variable is subject to missingness then the missingness pattern plays an important role in how we may proceed with the imputation task. With monotone missingness the incomplete variables may be imputed sequentially with the most observed variable being imputed first. This is done in the usual way by relating the incomplete outcome to other fully observed variables through an imputation model. Next, the second most observed variable is imputed by extending the imputation model to include the previously imputed variable as well. The process continues until the least observed variable has been imputed. The whole procedure is repeated m times to produce m imputed datasets.

The strength of this method is that different models may be employed with each imputed variable as appropriate. E.g. a logistic regression model may be used to impute one variable, a normal model to impute a second variable and predictive mean matching to impute a third variable. The monotone pattern implies that dependencies in the data are hierarchical: each incomplete variable is imputed by

modelling dependencies with more observed variables. Consider the monotone missingness pattern of the three variables shown in Figure 3.1. *Var3* is less observed than *Var2*, which is less observed than *Var1*. Unobserved values in *Var1* depend on neither of the other two variables. *Var2* may depend on *Var1*, but not on *Var3*. And *Var3* may depend on both *Var1* and *Var2*.

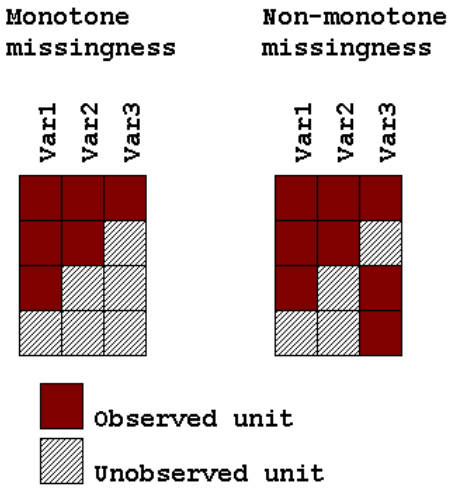


Figure 3.1. Monotone and non-monotone missingness patterns.

3.4.6 Non-monotone patterns

It is not uncommon in longitudinal studies from clinical research to have missingness that is caused by individual missed observations rather than drop-out. When missingness is intermittent in this way the resultant missingness pattern is non-monotone. Contrary to the situation just described there is no hierarchical dependency structure to guide the imputation process with non-monotone missing data. Using the example of the non-monotone pattern illustrated in Figure 3.1 it is clear that missing values in *Var2* may depend on *Var3* (as well as *Var1*), and equally that missing values in *Var3* can depend on *Var2*. Because of this interdependency between the incomplete variables we now have to model the variables jointly in some way rather than in succession.

Some authors (e.g. Robins & Gill, 1997) have pointed out that the MAR mechanism can seem incongruent with a non-monotone missingness pattern since missing data in

one variable may be predicted using data from other variables only when these other variables have been observed, i.e. the missingness mechanism that dictates the way in which the missing outcomes depend upon the observed outcomes changes with each observed missingness pattern in the dataset. Under a MAR assumption we do not have to model this mechanism. This is just as well because the number of non-monotone patterns grows exponentially with higher dimensional data. However even if we can avoid specifying these mechanisms we should still like to think that they exist in theory and that they are in some way plausible since the MAR assumption may otherwise not be justified.

3.4.6.1 Imputation based on a multivariate normal model using MCMC

When the missingness pattern is non-monotone an approach to the joint modelling of the multivariate data is needed. Clearly one multivariate distribution whose properties are well known is the multivariate normal distribution.

While the observed-data posterior distribution, $f(\theta| Y^O)$, is generally not a standard distribution from which realisations of θ are easily simulated, the full-data posterior for the parameters, $f(\theta| Y^O, Y^M)$, is typically more readily simulated but requires predictions for the missing values (Schafer, 1999). This circular problem can be solved iteratively by alternating between predicting the missing values and predicting the parameters. In each step, the predictions of either the parameters or the missing data are based on the latest prediction of the other. The iterations are set in motion by choosing some initial values of either the parameters or the missing data. The starting values may be simply a qualified guess or some estimate based on the available data. Eventually the sequence converges to the stationary distributions $f(\theta| Y^O)$, the observed-data posterior of the parameters, and $f(Y^M | Y^O)$, the predictive distribution of the missing data (Schafer, 1997, Chapter 3). The rate of convergence depends on the fraction of missing data and on the choice of starting values. Iterative algorithms of this sort are commonly referred to as Markov Chain Monte Carlo or MCMC.

3.4.6.2 Making imputations with MCMC

The iterative chain is started off by choosing a set of initial values for either the missing data or the parameters. Let $\theta^{(0)}$ denote some initial values for the parameters. These initial parameter estimates could for example be based on the complete cases. Next, the initial parameters are used to generate the first iteration $Y^{M(1)}$ of the missing values. These in turn are fed back into the posterior for the parameters to create $\theta^{(1)}$. The two steps

$$Y^{M(t+1)} \sim f(Y^M \mid Y^O, \theta^{(t)}) \tag{3.4}$$

$$\theta^{(t)} \sim f(\theta \mid Y^O, Y^{M(t)}) \tag{3.5}$$

are then repeated many times until convergence is reached. The Y^M from the last iteration are used to generate the first imputed dataset.

The sequence $(\theta^{(1)}, Y^{M(1)})$, $(\theta^{(2)}, Y^{M(2)})$, $(\theta^{(3)}, Y^{M(3)})$, …, $(\theta^{(t)}, Y^{M(t)})$ generated by cycling through (3.4) and (3.5) defines a Markov chain where, at each step in the chain, the simulated draws $(\theta^{(k)}, Y^{M(k)})$ are statistically dependent on the draws $(\theta^{(k-1)}, Y^{M(k-1)})$ from the previous step (Schafer, 1999). MI requires *independent* draws from the stationary distributions to produce *m* separate datasets. These may all be obtained from the same chain by allowing for a large number of iterations between each draw to ensure that the correlation between consecutive draws is negligible. Alternatively separate chains may be used for each of the *m* imputed datasets.

3.4.6.3 When is convergence obtained?

Although convergence to the stationary distributions can be demonstrated to have been obtained in some simple examples it is notoriously difficult to ascertain exactly when this happens in general. Various techniques exist for assessing when convergence has occurred. These include inspecting time-series plots of the simulated draws from consecutive iterations of the MCMC chain, running parallel chains initiated with different starting values and calculating measures of autocorrelation within chains (Schafer, 1997, Section 4.4).

3.4.6.4 Why the multivariate normal distribution?

In principle the above modelling method could be applied with other multidimensional probability distributions other than the multivariate normal distribution. In practice such alternative multivariate distributions are a challenge because we would need to specify the full joint distribution to generate the random draws. This is in contrast to MI methods using Chained Equations (see below) that do not require the specification of a genuine joint distribution.

Interestingly, some good results have been obtained with MI based on the multivariate normal model even for variables that are clearly not normally distributed. Success has been reported not only with ordered categorical, but even binary data (Schafer, 1997, Section 5.1). MI based on the multivariate normal model can perform well when the fraction of missing data is small and when the imputed variables are not too skewed.

Molenberghs & Kenward (2007) point out that the sequential regression method and the MCMC method should yield almost identical results when the data to be modelled are from a genuine multivariate normal distribution and the missingness pattern is monotone. However there may still be small discrepancies depending on the choice of prior for the parameters and when the sample size is too small to rely on asymptotic normality.

3.4.7 MI using chained equations (MICE)

More recently an alternative MI method for the imputation of multivariate non-normal data has been put forward. The method known as MI using Chained Equations (MICE) or MI using Fully Conditional Specification (van Buuren, 2007) differs from the multivariate normal imputation previously described in that it does not require the specification of a joint multivariate distribution for the variables to be imputed.

The use of MICE is particularly useful with non-monotone missing data in datasets involving different types of variables. For example, using MICE it is straightforward to impute inter-dependent binary, ordinal, categorical and quantitative variables within the same dataset. A separate univariate model is specified for each variable to be imputed such that an appropriate model can be chosen for the particular type of data being imputed. E.g. a logistic model can be used for binary variables, an ordinal regression for ordered categorical variables etc. MICE then uses an iterative algorithm to model the conditional distribution of each incomplete variable conditional on all the other variables in the model. Starting with a set of initial imputations for the missing data, the conditional models are processed in sequence and the imputations updated at each iteration. By continually conditioning on the most recently updated version of the predictor variables the imputed variables eventually converge to a stationary distribution which, it is hoped, correspond to a true underlying joint distribution of the incomplete variables.

In fact it is not guaranteed that the limiting distribution of the imputations obtained in this manner will converge to some theoretical joint distribution (Kenward & Carpenter, 2007). However, despite the lack of a theoretical basis this method has produced good results in practical applications and simulation studies (e.g. Lee & Carlin, 2010). van Buuren (2007) concludes that MICE provides: "a useful and easily applied flexible alternative to JM [joint modelling] when no convenient and realistic joint distribution can be specified."

3.4.8 Is MI just making up data?

MI is a powerful method for gaining inference with incomplete data through efficient use of the available data. Unlike the direct likelihood method, MI does rely on stochastic imputations of the missing data. Some find this idea unsettling and may dismiss MI as a procedure for making up data where none was observed. Certainly the gain in efficiency from using MI over a simple analysis of the available data may seem to suggest that we are getting something for nothing. However this efficiency gain is just as real as that achieved through a likelihood-based analysis and is simply the result of exploiting the already existing dependency structures in the data. MI

provides a flexible alternative when direct likelihood methods are impractical or demand the specification of unfeasibly complicated models. In fact, MI and the method of direct likelihood are asymptotically equivalent in the sense that one would obtain virtually identical results using either approach when the number of imputations *m* is large.

There are assumptions behind any analysis with missing data. A complete-case or available-case analysis may be less controversial with some researchers than an analysis based on MI as these types of analysis do not force us to make explicit assumptions about the missing values. In fact a complete-case analysis relies on much stronger (but implicit) assumptions about the unobserved values than does a MI analysis.

3.5 The fully Bayesian approach

In a fully Bayesian analysis missing data are treated as unknown quantities on a par with unknown parameters (Lunn et al., 2013). Just as an unknown parameter is estimated in terms of a posterior distribution, a posterior-predictive distribution, f($\tilde{y}$ |y) for the missing data can be derived from the data likelihood and prior specifications for the parameters. The predictive distribution can then be used to simulate realisations of the missing values as is the case with MI. However deriving the posterior-predictive distribution for the missing data is not a requirement for obtaining valid posterior estimates of the parameters. Posterior distributions for the model parameters can be readily obtained as long as the data likelihood has been fully specified along with the relevant prior distributions. This method deviates from the direct likelihood method only to the extent that the prior distributions contribute to the Bayesian analysis. The two methods are essentially equivalent when the prior distributions are flat relative to the data likelihood. There are obviously also very close ties between the fully Bayesian approach and multiple imputation methods since the latter rely on the Bayesian predictive distribution to simulate realisations of the missing data, the *m* imputed datasets. Thus MI and the fully Bayesian approach are essentially equivalent when *m* approaches infinity. The use of Bayesian methods for the analysis of incomplete data will be further explored in Chapters 6, 7 and 8.

3.6 Informative missingness

In the present chapter we have primarily discussed methods for handling data under the assumption that the data were MAR. However in practice it is rarely possible to rule out the possibility that the data are MNAR.

When that is the case, missingness is said to be *informative* because the very fact that an observation is missing has a direct bearing on the unobserved value itself. The missingness process is then no longer ignorable and it is necessary to model the full joint data-likelihood,

$$L_{full} = f(Y^O, Y^M, R \mid \theta, \psi)$$

In section 3.2 we showed a factorisation of this function known as the selection model,

$$f(R \mid Y^O, Y^M, \psi)\, f(Y^O, Y^M \mid \theta)$$

Another common factorisation is

$$f(Y^O, Y^M \mid R, \theta)\, f(R \mid \psi)$$

This is known as the pattern-mixture model (PMM) factorisation. (Molenberghs & Kenward (2007, Chapters 17 and 24) also give details of a third family of models, shared-parameter models, that allow the joint distribution for Y and R to be specified in terms of shared random effects or latent variables. These model are less common and, although interesting and with the offer of added flexibility, we are unable to consider these further within the scope of the current project).

Whether fitting a selection model or a PMM, when the missingness process is non-ignorable it is necessary to specify the relationship between R and Y^M. This is done

either through $f(R \mid Y^O, Y^M, \psi)$ in the selection model or through $f(Y^O, Y^M \mid R, \theta)$ in the pattern-mixture model.

The nature of this relationship cannot be estimated from the incomplete data itself. The model specification is therefore often the result of untestable assumptions based on the context that gave rise to the data, and perhaps some external evidence.

Because of the inherent uncertainty associated with analysis of MNAR data it is not unusual to base a main analysis around the MAR assumption, but then refit the model under several different MNAR scenarios as a means of testing the sensitivity of the main findings to deviations from the MAR assumption. For example, using a selection model it may be reasonable to specify the non-response mechanism, conditional on the unobserved data, such that the deviation from the MAR assumption is quantified in a single parameter. Letting π_i denote the probability that Y_i is missing we can specify a model for π_i

$$logit(\pi_i)=\alpha+\beta Y_i$$

such that $\beta=0$ is consistent with the MAR assumption while any non-zero value for β represents a deviation from the MAR assumption. β cannot itself be estimated from the data. α represents the level of missingness in Y_i when $Y_i=0$ and is estimated from the data once β is fixed. The model is evaluated under a range of plausible values for β and the impact on the main findings assessed.

In the present chapter we have presented the principled approaches of direct likelihood, MI and the fully Bayesian approach primarily in the context of ignorable missingness. Each of these methods can be applied when the missingness process is non-ignorable. As we have seen this requires that the full joint likelihood for the response and non-response mechanism be fitted. Such likelihood functions tend to be intractable and require numerical optimisation techniques. MNAR models are not easily accommodated in most standard statistical software packages, although it is possible to model the complex likelihood functions that arise from fitting such

models using MCMC methods, for example as implemented in WinBUGS. Analysis of MNAR data will be explored further in chapters 6, 7 and 8.

4 REGRESSION TO THE MEAN

The Symptom Monitoring Service (SMS) was introduced briefly in section 1.2 and serves as an example of a clinical service that routinely collects patient outcomes over time at intervals determined by the timing of patients' appointments. Patients were asked to complete the distress symptom screening questionnaire in the waiting areas of screened oncology outpatient clinics, either on paper or on touch screen computers, whilst waiting for their consultation.

The SMS questionnaire enquired about symptoms of psychological distress and it was therefore important to assess whether a medical clinic is a suitable setting in which to screen patients for distress. It is possible that patients' distress scores were affected by the clinic surroundings, and in anticipation of the imminent appointment. If it were the case that distress scores were transiently inflated due to the clinic setting, it would question the clinical usefulness of the ratings and the validity of the screening approach in general.

The POD Study, which is the subject of Chapter 5, examines the persistence and development over time of symptoms in cancer outpatients who were identified with significant distress symptoms in clinic. It is relevant therefore whether a high score in clinic is a good indicator of patients' true underlying distress status, and whether the clinic scores are comparable to scores obtained over the telephone from patients when in their own homes.

In addressing these questions this chapter will focus on the nature of repeated distress scores. We will consider how scores are correlated and how the correlation may be used to model changes over time including regression to the mean, and how this relates to the idea of persistence.

4.1 Background

The Symptom Monitoring Service (SMS) operated in selected National Health Service oncology outpatient clinics in Scotland, UK, between May 2008 and August 2011. Patients attending for appointments at a screened clinic were approached in the

clinic waiting area and asked to complete a questionnaire which enquired about physical and psychological symptoms. The questionnaire answers were used to help the oncologists address issues that were of concern to the patients, and also to help identify patients with major depression for subsequent eligibility assessment for enrolment into the SMaRT Oncology-2 and -3 trials.

Prior to each clinic the SMS received a complete list of names of patients scheduled for appointments. The service aimed to screen all patients who attended, although a small proportion of patients did not complete screening because they were missed (typically because they were taken straight for their appointments before the SMS could approach them), were excluded from the screening service on medical grounds or refused to complete the symptom screening questionnaire.

There were three parts to the SMS questionnaire. The first part consisted of the Hospital Anxiety and Depression Scale (HADS) (Zigmond & Snaith, 1983), which asks patients about symptoms of emotional distress. It consists of two subscales: an anxiety and a depression subscale each containing 7 items scored on a 0 to 3 scale, resulting in a total HADS score ranging from 0 (no distress) to 42 (maximal distress). The second part consisted of the five questions from the EQ-5D (The EuroQol Group, 1990). Finally in the third part, patients were asked to rate on a 0-to-10 scale how bothered they had been over the last week by each of the following symptoms: pain, fatigue, disturbed sleep and nausea or vomiting.

To identify patients likely to suffer from depression, the SMS sought to follow up on everyone who was identified with symptoms of significant distress. Scoring 15 or more on the HADS has been shown to provide a good indication of depressive or anxiety disorder (Walker et al., 2007). Patients who scored 15 or more on the HADS in clinic were therefore telephoned one or two weeks later and asked to complete the part of the SCID (Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders (DSM-IV); First et al., 1999) that relates to depression. Because the HADS scores that were collected in clinic were used to identify patients likely to suffer from depression, it was important to establish if these scores provided a good

picture of patients' psychological well-being. Might the service be labelling high scoring patients as distressed when actually their scores were artificially inflated due to the potentially stressful circumstances? The implications would be that too many patients were interviewed for depression during the subsequent follow-up telephone call causing unnecessary inconvenience to patients and wasting the resources of the service.

To address the issue, the service conducted an audit during the period from March to April 2009. Patients who scored high on the HADS in clinic (score ≥ 15) during this period were extraordinarily asked to complete the HADS a second time at the beginning of the telephone interview for depression routinely carried out in the high scorers approximately seven days after the clinic appointment. Patients' scores on the second HADS were then compared to the scores obtained in clinic.

Specifically, the purpose of the audit was to (a) determine the number of patients who scored 15 or above on the HADS in clinic who no longer scored above 15 when assessed at home, and (b) determine the mean change in scores over the two assessments.

4.2 A challenging design

Ideally the questions posed above would be addressed using a suitable test-retest design in which all patients were followed up regardless of their clinic HADS score. This would allow for a direct interpretation of the mean scores and observed proportions at the two time points.

The SMS screened a large number of patients each day. Around a fifth of those who were screened scored high on the HADS and needed further assessed over the telephone. To follow up on everyone, the service would have had to interview five times as many patients. As this was not a feasible option, only patients who scored high in clinic were asked to complete the scale a second time as part of the audit.

However, unless there is perfect correlation between repeated HADS scores we would expect to see some regression to the mean. This means that at least some of the observed change will not reflect a real difference in scores over the two assessments.

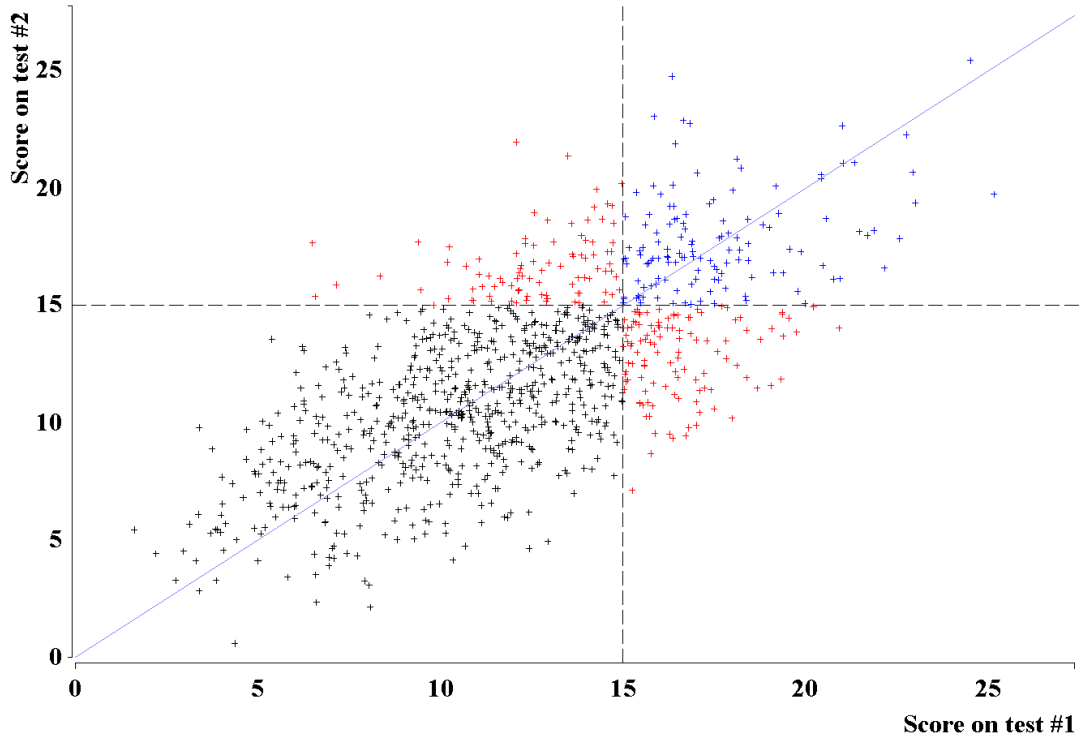


Figure 4.1. Simulation of 1000 test-retest patient scores from a stationary distribution. The cut point separates low scores from high scorers on the two occasions. Patients who scored on different sides of the threshold on the two tests are indicated in red. The diagonal line is the line of perfect agreement.

Senn (2009) used an illustration similar to that in Figure 4.1 to illustrate the artificial effect induced by regression to the mean. Patients who score high will on average score lower on reassessment. Of the patients who score above the threshold on the first assessment (right side of the vertical dashed line) a considerable proportion proceed to score below the threshold on the second assessment (bottom right corner) despite there being no overall change in scores between the two assessments.

4.3 Why might patients score differently on reassessment?

Conceptually we may think of an observed HADS score as resulting from three constituent parts: Firstly, each patient can be thought of as having a hidden true score

representing their actual distress level. Secondly the observed score is affected by factors influenced by the patient's immediate environment (e.g. was the questionnaire read out to the patient over the telephone, or handed to them on paper). Finally, the score is also affected by random measurement error.

All three contributing factors can change over time. Indeed, patients who took part in the HADS audit were reassessed approximately one week after the clinic assessment; it would be naïve to assume that the underlying distress levels of individual patients were constant over this interval. In the immediate environment things were different too. The HADS was completed over the telephone from the patient's own home, and not on paper or touch-screen computer in the clinic. Finally, random error also has the potential to cause two very different scores. The possible causes behind patients in the HADS audit scoring differently at home compared to when in clinic are listed in Table 4.1.

Table 4.1. Reasons for scoring differently on reassessment

1.	Patients' true underlying scores could have changed since the first assessment.
2.	The questionnaires were administered over the telephone on the second occasion. This might have caused patients to score the instrument differently.
3.	Patients might have been anxious about the outcome of the upcoming appointment when assessed in clinic. This could have caused them to score higher in clinic.
4.	Random error in the measurements (not including regression to the mean) may have resulted in spurious change.
5.	The findings may have been confounded by regression to the mean due to the design.

Considering each of these causes we note firstly that while individual patients may experience changes in their underlying, true distress levels over a period of seven days there is no reason to believe that the sample as a whole should have changed. Secondly, it does not seem very likely that patient scores were considerably different when collected over the telephone as opposed to on paper or touch-screen computer.

Indeed Pinto-Meza et al. (2005) examined the method of administration of the PHQ-9, a similar instrument to the HADS, and found that there was good agreement between self-administered and telephone-administered questionnaires. Lastly, the fluctuations caused by random error will average out with increasing *n*, the number of patients sampled, since the mean of the error term is zero. Practically therefore, the only plausible reasons for an overall change in the scores are the different setting (clinic versus home) and regression to the mean.

4.4 Regression to the mean

Having established that any observed change in scores on reassessment is likely to be due, at least in part, to regression to the mean (RTM), can we say anything more about the likely size of this effect?

The RTM effect depends on the strength of the correlation in the data. If there is no correlation between repeated scores, a sort of memory-less sequence, then the RTM effect will be at its greatest. The distribution of individual patients' scores will be the same as that of scores in the overall population. If there is perfect correlation then there is no measurement error, and repeated scores from the same patient will be exactly equal, i.e. there is no RTM effect.

Crucial to the question is therefore what we might expect the correlation to be between repeated scores on the HADS. Distress is a dynamic variable, so the answer presumably depends on the time between the assessments. Perhaps we should expect a near-perfect correlation in assessments obtained just after one another, while, as time goes on, symptoms change and correlations become weaker. Although we might expect a weakening correlation over time it seems unlikely that this would ever approach zero: Even after many, many months there might still be some subject effect due to invariable personality or mood traits. Further on in section 4.7 we shall focus on modelling this correlation over time in order to estimate the correlation of scores obtained seven days apart.

The effect of regression to the mean has been well known for a long time. Galton (1886) described the phenomenon in his article "Regression towards mediocrity in hereditary stature" examining the relationship between the heights of adult children and their parents. Nonetheless, researchers unaware of the pitfalls of RTM continue to misinterpret their findings from time to time. Numerous papers have been published on the topic warning researchers of the untoward effect, recommending study designs to alleviate the effect, and advising on how to analyse data appropriately in the presence of RTM (e.g. Davis, 1976; Das & Mulder, 1983; Beath & Dobson, 1991; Senn, 2007; Barnett, van der Pols & Dobson, 2005).

We know that changes observed within the HADS audit test-retest design would have been affected by regression to the mean, and we wish to estimate the size of this effect. The effect can be described as the expected difference between a pair of repeated HADS scores, conditional on the first score being equal to, or more than, 15:

$$E[H_1 - H_2 \mid H_1 \geq 15]$$

Above, H_1 and H_2 are random variables denoting the HADS scores at time one and two respectively. In sections 4.6 – 4.9 we will apply the approach developed by Das & Mulder (1983) to estimate this quantity. For now however we will focus on the scores actually observed in the HADS audit.

4.5 Results

During the audit period 395 patients were identified as high scorers in clinic. Eighty three percent of these (329 patients) were eligible for a follow-up assessment and 218 patients (66% of those eligible) were successfully assessed a second time and their data included in the analysis. Reasons for exclusion are shown in Figure 4.2. The main reason for exclusion was that the patient could not be telephoned by the service within the reassessment time window used for our analysis.

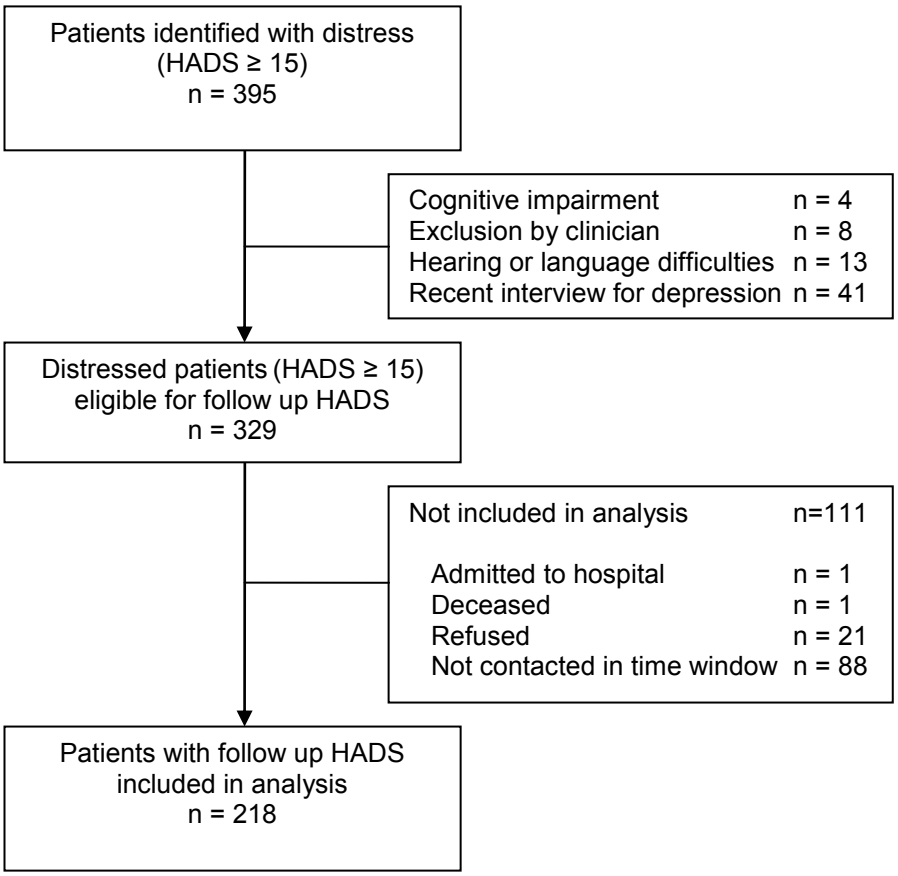


Figure 4.2. Derivation of sample of patient data used in the analysis.

The patients whose data were included in the analysis had attended clinics that specialised in breast (83), colorectal (18), gastrointestinal (15), gynae (33), lung (47), sarcoma (4), urology (14) and miscellaneous cancers (4). Seventy three percent (159/218) of the patients were female and the median age was 61 years (interquartile range: 53 to 70 years). The HADS scores, age and sex distributions of the sample included were similar to those of eligible patients who were not included in the analysis (Table 4.2).

The median interval between the clinic and follow up assessments was six days (interquartile range: 5 to 8 days). The distributions of HADS scores when: (a) patients were assessed in clinic and (b) when followed up approximately one week later at home are shown in Figure 4.3.

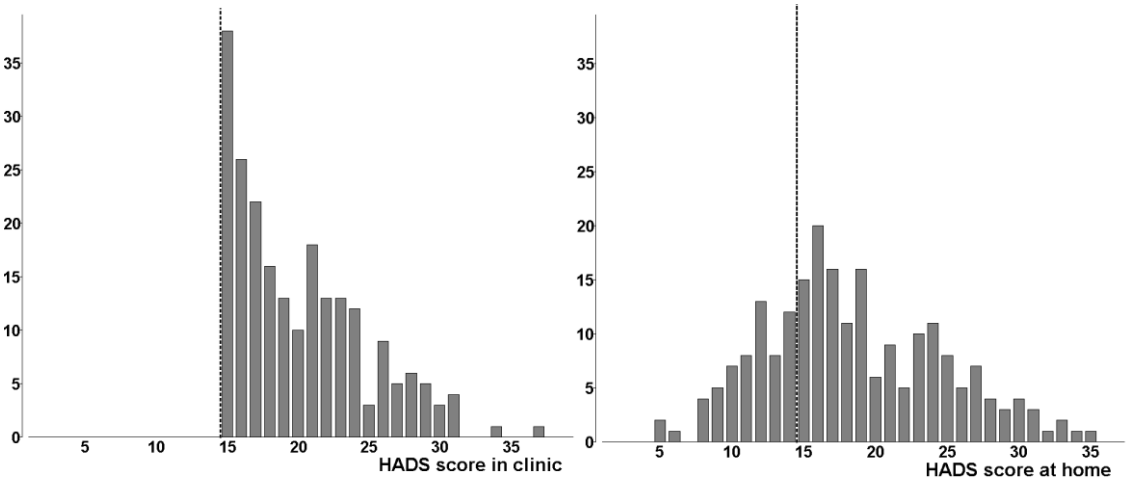
Table 4.2. Characteristics of the analysed sample compared with those of eligible patients not included.

	Eligible patients included for analysis n = 218	**Eligible patients not included for analysis** N=111	**P-value**[a]
Age in years			0.954
Mean (SD)	61.4 (11.5)	61.3 (12.2)	
Median (range)	61.4 (25.3 to 87.7)	62.5 (28.7 to 89.8)	
Age categories:			0.736
≤50	38 (17%)	23 (21%)	
51-60	67 (31%)	28 (25%)	
61-70	66 (30%)	35 (32%)	
≥ 71	47 (22%)	25 (23%)	
Gender			0.396
Male	59 (27%)	35 (32%)	
Female	159 (73%)	76 (68%)	
HADS scores			0.356
Mean (SD)	20.1 (4.7)	20.6 (4.8)	
Median (range)	19 (15 to 37)	19 (15 to 34)	
HADS score categories			0.604
15-19	115 (53%)	59 (53%)	
20-24	66 (30%)	29 (26%)	
≥ 25	37 (17%)	23 (21%)	

[a] Age in years and HADS scores were compared using Wilcoxon rank sum test. The three other p-values were from chi-square tests.

Figure 4.4 shows the change in HADS score from clinic to follow-up plotted against initial HADS score in clinic. There was considerable variability in the change scores with some patients scoring much higher and some much lower on reassessment. Most patients whose scores fell below 15 at follow up scored only slightly above 15 in clinic.

Almost three quarters (72.5%; 158/218) of patients remained high scorers on the follow up assessment (95% CI: 66.6 to 78.4%). Most patients who were no longer high scorers continued to have elevated scores (between 10 and 14). The mean score at follow-up was 18.4, a reduction of 1.74 units (95% CI from 1.09 to 2.39) from the clinic visit. The reduction in the anxiety subscale score was 1.26 units (95% CI: 0.84 to 1.67) while the depression subscale dropped by just 0.48 units (95% CI: 0.12 to 0.85). The difference in the reduction between the subscales was statistically significant ($p<0.001$).



(a) (b)

Figure 4.3. HADS scores of patients (n=218) in the study sample (a) when assessed in clinic and (b) when reassessed at home.

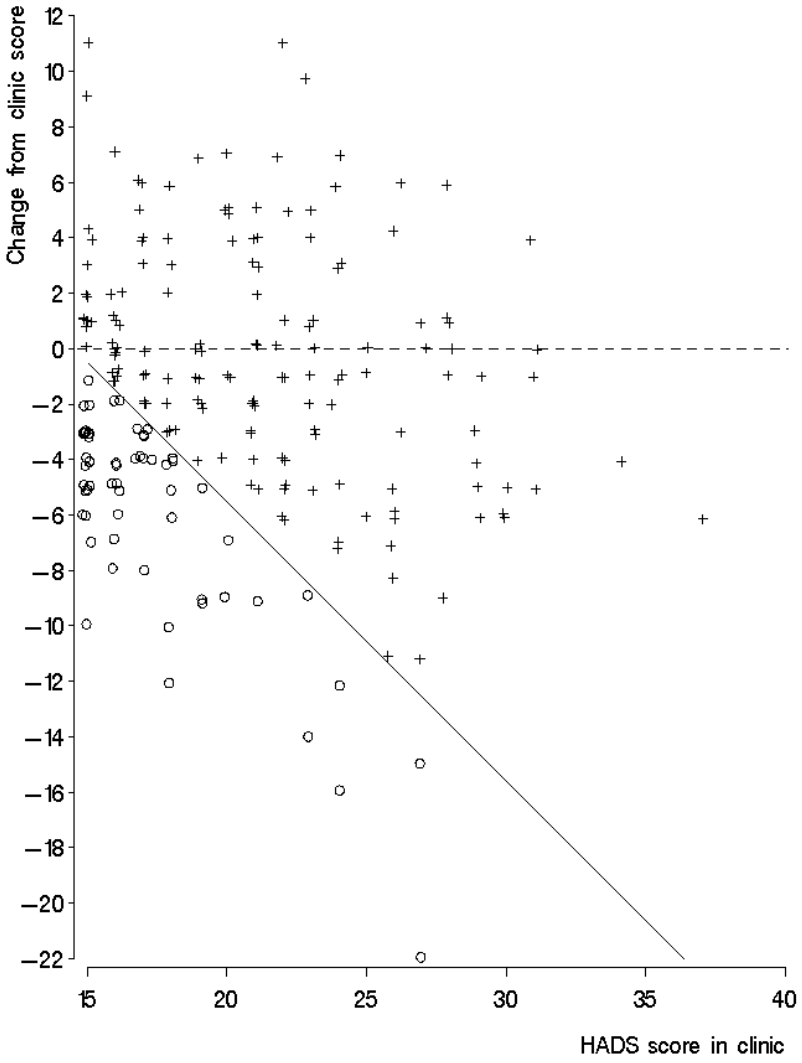


Figure 4.4. Change in HADS total score from clinic to follow up plotted against initial HADS score in clinic. Circles indicate patients whose reassessment score fell below 15. Patients plotted above the dashed line had a higher HADS score on reassessment while those below the line had a lower score. A degree of 'jitter' has been applied to separate out overlapping data points.

4.6 Estimating the RTM effect

When completing the HADS at home patients scored on average 1.74 points lower on the total scale than they had done a week earlier in the clinic, with a larger reduction on the anxiety subscale. Between a quarter and a third of patients scored below the distress threshold of 15 on the follow up HADS. Some of this change may reflect a real difference in scores on the two occasions, presumably because of the change in setting. On the other hand, we would expect part of this change to be an artefact caused by regression to the mean. In the following sections we will attempt to quantify this effect.

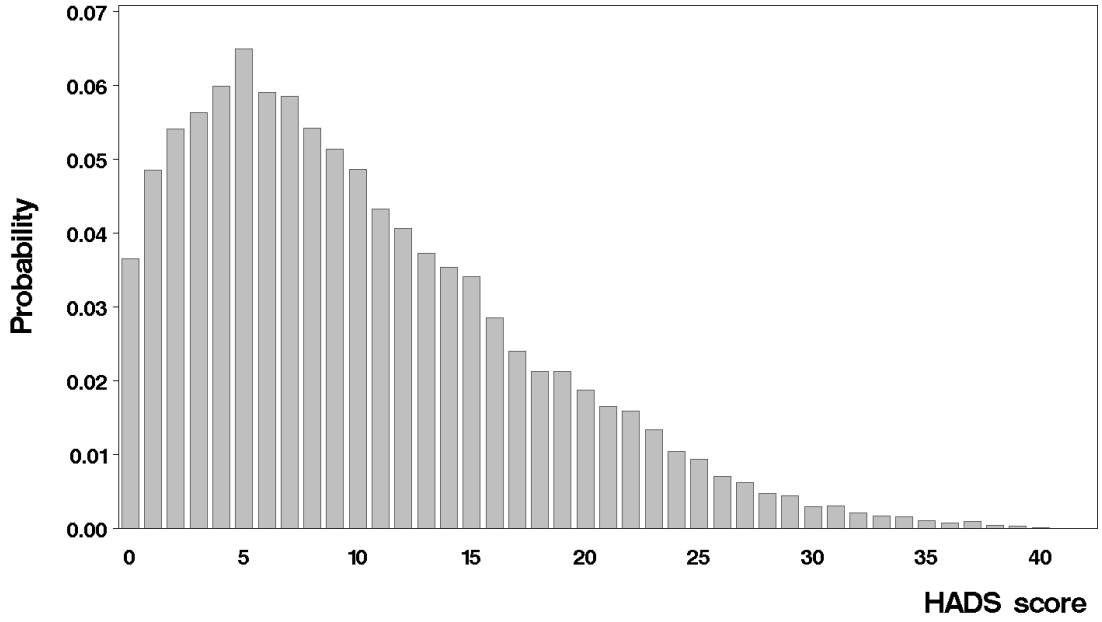


Figure 4.5. Distribution of outpatient HADS scores.

The distribution of HADS scores as collected by the Symptom Monitoring Service from the outpatient cancer population is presented in Figure 4.5. The scores are treated as random variables that are subject to variance terms. Some of the variance comes from inter-subject variability: patients are different and will score according to their individual mood characteristics. The remainder of the variance will be attributed to intra-subject variability, akin to random variation or measurement error within a patient's scores.

The total variance, σ_t^2, can therefore be written: $\sigma_t^2 = \sigma_s^2 + \sigma_e^2$. Here, σ_s^2 denotes the variance arising from heterogeneity between subjects, while σ_e^2 is the random error. σ_s^2 is also the covariance between repeated measurements within the same subject, and $\rho = \sigma_s^2 / \sigma_t^2$ is the correlation.

The HADS score, *H*, can then be thought of as the sum of the random variable *S*, denoting the patients' true score, and *e*, the random error. *S* is distributed according to some density function *f* with variance σ_s^2. We will assume that the random errors are distributed according to $N(0, \sigma_e^2)$.

We wish to estimate the expected difference between a pair of repeated HADS scores, H_1 and H_2, conditional on the first score being equal to, or more than, 15, $E[H_1 - H_2 \mid H_1 \geq 15]$.

For a continuous *H* it can be shown that

$$E[H_1 - H_2 \mid H_1 > h_c] = (1-\rho)\,\sigma_t^2 \,\frac{g(h_c)}{1 - G(h_c)}$$

Above $g(h_c)$ is the probability density function for *H* evaluated at h_c and $G(h_c)$ is the corresponding cumulative distribution function. This approach by Das & Mulder (1983) allows for an arbitrary *g*, although the errors are assumed normally distributed.

From the large sample of scores collected by the Symptom Monitoring Service we may obtain estimates of $g(h_c)$ and $G(h_c)$. Estimating these quantities is the topic of section 4.8.

Given appropriate values for $g(h_c)$ and $G(h_c)$ the effect size depends on $(1-\rho)\,\sigma_t^2$. The RTM effect is proportional to $(1-\rho)$. A central task is therefore estimating ρ, or equivalently the variance and covariance, of scores approximately one week apart.

4.7 Estimating correlation parameters

Only a very small proportion of patients screened by the Symptom Monitoring Service were followed up more frequently than once a month. Hence there is very little data available on scores collected just one week apart. Even if the service had followed-up sufficiently many patients at this short an interval, such patients would likely differ from patients in the audit on disease and demographic characteristics. Estimates of the correlation, and indeed of the RTM effect, would not necessarily generalise to outpatients in general.

Instead the SMS data will be used to model the variance-covariance parameters as a function of time between repeat assessments. Once the model has been fitted it can be used to evaluate the parameter values at just seven days.

In the first instance we will assume that the covariance of repeated scores is independent of the time between the assessments, i.e. that the covariance is constant over time.

The following linear model (Model 1) was fitted to the data

$$Y_i = X_i\beta + e_i$$

Here Y_i denotes a vector of length n_i of HADS scores belonging to patient i.

$$X_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}$$

is the design matrix with t_{ij} denoting the time (in weeks) of the jth visit for patient i and $\beta = [\beta_1\ \beta_2]'$ is the corresponding vector of regression coefficients.

$$e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}$$

is the vector of random errors. The e_{ij}'s are distributed according to the multivariate normal distribution with *0* mean vector and covariance matrix Σ with all diagonal elements of Σ equal to $\sigma_s^2 + \sigma_e^2$ and all off-diagonal elements equal to σ_s^2.

The model parameters were fitted by (restricted) maximum likelihood estimation (REML) using the Mixed Procedure in the SAS software.

Fitting this model the variance was estimated to be $\sigma_t^2 = 52$, the covariance $\sigma_s^2 = 37$, and therefore the correlation between repeated measurements within the same subject $\rho = 0.71$.

It seems unlikely that the covariance, or equivalently the correlation, should be constant over time. A time-dependent covariance pattern was therefore fitted instead (Model 2) to allow correlations to deteriorate over time using the following functional form:

$$\sigma^2 \mathrm{P}^{(t_{ij} - t_{ik})}$$

Here $(t_{ij} - t_{ik})$ is the time (in weeks) between the *j*th and the *k*th measurement. This so-called exponential covariance pattern will ensure perfect correlation when $t_{ij} = t_{ik}$, but also imposes an increasingly weak correlation when assessments are further separated in time. Using this pattern for the covariance yielded the following estimates: $\sigma^2 = 53$ and $P = 0.95$. According to this model the correlation between two measurements one week apart was therefore *0.95*. The correlations quickly weakened to *0.85* after four weeks, *0.60* after three months and *0.30* after half a year. The correlation coefficient between two measurements one year apart was just *0.08* according to this model. As discussed in section 4.4 it may be reasonable to allow for

some dependency to remain between measurements taken from the same subject regardless of the time between the assessments. This feature will be accommodated in Model 3.

Before introducing the final model, we note that the compound symmetry from the covariance matrix in Model 1 can equally be achieved by fitting a linear model with a random intercept.

$$Y_i = X_i\beta + b_i + e_i$$

Here $b_i \sim N(0, \sigma_s^2)$ is a subject-specific random offset imposing dependency between repeated measurements within the same subject. $e_i = [e_{i1}, e_{i2}, \ldots, e_{in_i}]'$ is the vector of independent, identically distributed measurement errors with $e_{ij} \sim N(0, \sigma_e^2)$. With this model the covariance matrix for the responses is $Cov(Y_i) = V$ with all diagonal elements of V equal to $\sigma_s^2 + \sigma_e^2$ and all off-diagonal elements equal to σ_s^2 exactly as in the first model.

We wish to induce a time-dependency in the correlations without forcing a near-zero correlation when measurements are far apart or a perfect correlation when measurements are close in time. This may be achieved by relaxing the assumption of independence among the e_{ij}'s in the random intercept model. Instead e_{ij} will be modelled according to the multivariate normal distribution with mean *0* and covariance matrix R with all diagonal entries equal to $\sigma_e^2 + \sigma^2 P^{(t_{ij} - t_{ik})}$ and all off-diagonal entries equal to $\sigma^2 P^{(t_{ij} - t_{ik})}$. The four covariance parameters fitted using this model (Model 3) were $\sigma_s^2 = 34$, $\sigma_e^2 = 7$, $\sigma^2 = 10$ and $P = 0.86$. The variance

$$Var(Y_{ij}) = \sigma_e^2 + \sigma_s^2 + \sigma^2 = 51.8$$

is assumed to be equal at all time points. The covariance between repeated measurements

$$Cov(Y_{ij}, Y_{ik}) = \sigma_s^2 + \sigma^2 \mathrm{P}^{(t_{ij} - t_{ik})}$$

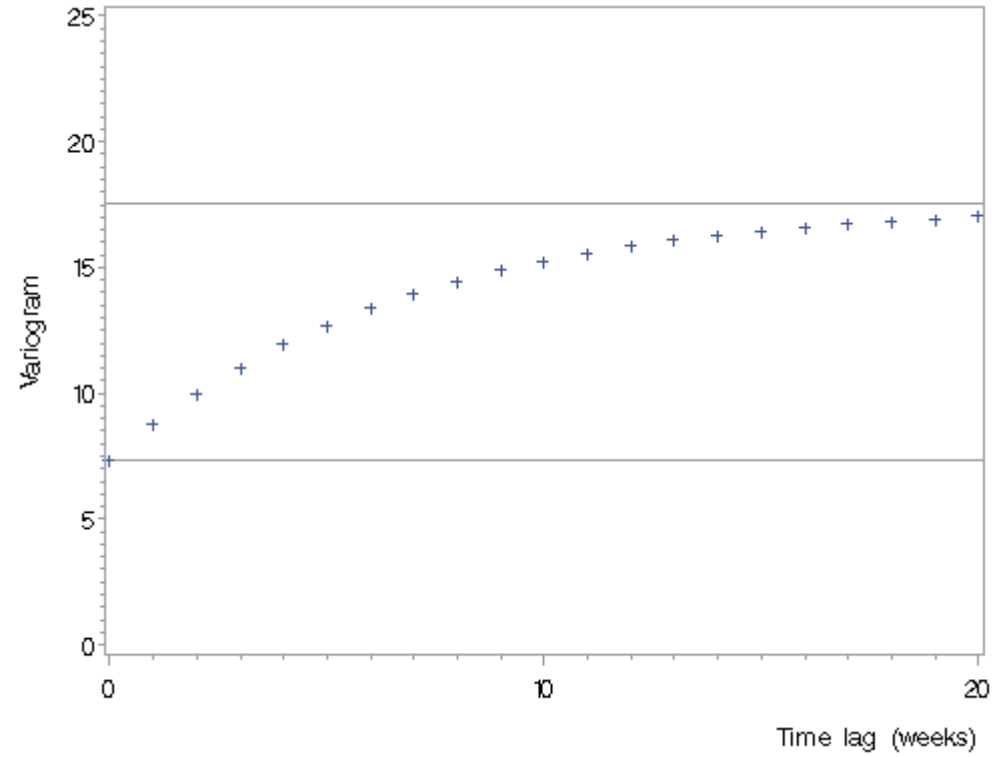
is time-dependent but bounded by $\sigma_s^2 + \sigma^2 = 44$ when $t_{ij} - t_{ik} = 0$ and $\sigma_s^2 = 34$ when $t_{ij} - t_{ik} = \infty$. This means that ρ, the correlation between repeated measurements, is between *0.66* and *0.86* depending on the separation in time between the assessments. Evaluated at $t_{ij} - t_{ik} = 1$ week the correlation is $\rho = 0.83$.

We opted not to base the analysis on the observed correlation between pairs of measurements obtained seven days apart for fear that patients with such abnormal clinic visit patterns would not be representative of the sample in general (and due to the limited number of such observed pairs). Nonetheless, having modelled the correlation it is still of interest to compare the resultant estimate with the observed correlation between measurements obtained seven days apart because good agreement between the two should increase our confidence that the modelled estimate is a reasonable one, more so than if the two differed markedly. (Of course there is still the possibility that both estimates could be wrong albeit for different reasons.) The *observed* correlation coefficient between 64 pairs of measurements obtained seven days apart in the SMS data was *0.86* (95% CI: *0.78* to *0.91*). The modelled estimate therefore appears to be in good agreement with the observed correlation in the data.

The -2 x log-likelihood statistics for Models 1, 2 and 3 above were *90727*, *92023* and *90548* respectively. Model 2 is a special case of Model 3 with $\sigma_e^2 = \sigma_s^2 = 0$. Model 1 is also a special case of Model 3 with $\sigma^2 \mathrm{P}^{(t_{ij} - t_{ik})} = 0$. Likelihood ratio tests comparing Model 3 to Model 1 ($G^2 = 180$; $df = 2$; $p < 0.001$) and to Model 2 ($G^2 = 1475$; $df = 2$; $p < 0.001$) provide evidence that Model 3 provided in a better fit. The Akaike Information Criterion (AIC) values were *90731.4*, *92027.0* and *90555.7* for the three models. Again Model 3 provides the better fit also according to this criterion.

Extensive exploratory analysis of the correlation structures in the data suggested a slow decay of the strength of correlation between pairs of observations with increasing time separation. The covariance structure fitted under Model 3 allows for a correlation between pairs of observations that decreases exponentially with the separation in time between the points. The exponential correlation model is immediately obvious when using the parameterisation $P = e^{-\theta}$. This, when coupled with σ_e^2 and σ_s^2, provides a flexible model that allows for a wide range of covariance patterns. We could have justified this model by testing it against an even more flexible model with further covariance parameters, however as is apparent from the variograms (Diggle et al., 2002) presented in Figure 4.6, the covariance structure fitted in Model 3 provides a reasonable fit to the covariance structure in the sample.

(a)



(b)

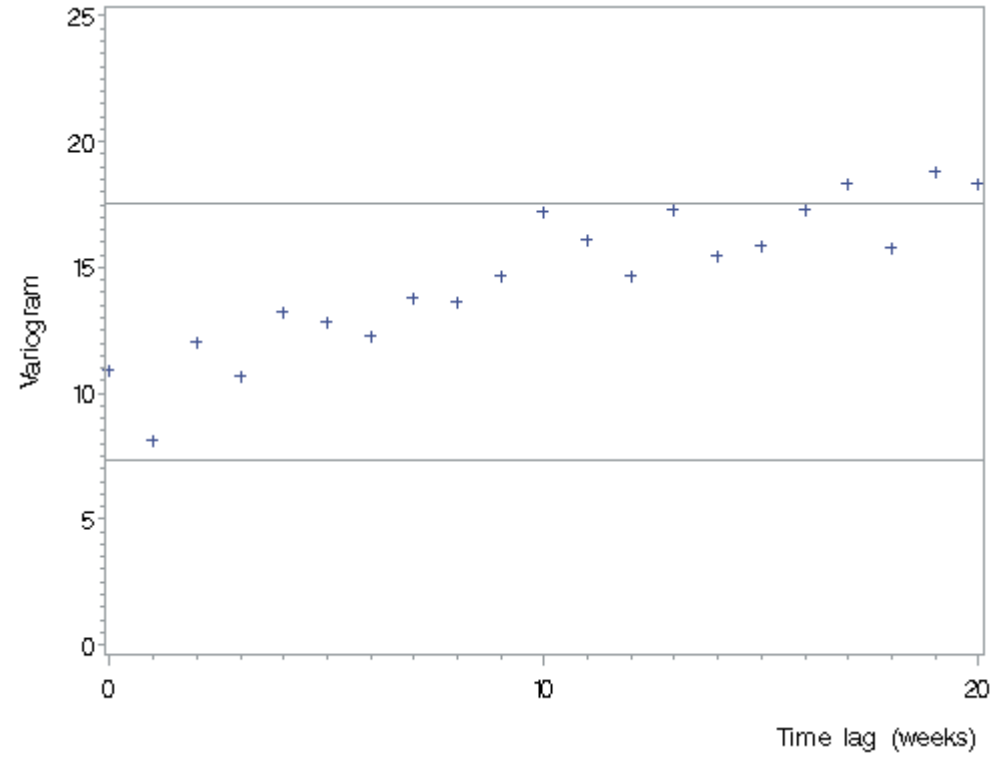


Figure 4.6. Diggle's variogram (a) under Model 3, and (b) derived from the sample residuals for time lags ≤ 20 weeks. The lower and upper horizontal lines indicate the model estimates of σ_e^2 and $\sigma_e^2+\sigma^2$ respectively.

4.8 The continuity problem

In the best fitting model for the covariance above, the covariance of two measurements, Y_1 and Y_2, obtained one week apart was estimated to be $Cov(Y_1,Y_2) = 43$, accounting for $\rho = 43.0/51.8 = 0.83$ of the total variance. We now need to determine the values of $g(h_c)$ and $G(h_c)$, the density and the cumulative distribution functions evaluated at h_c, the cut-off for significant distress.

Determining these quantities requires knowledge of the overall distribution of HADS scores in the outpatient population. While we may not know the algebraic form of the probability density function, we can obtain the empirical distribution from the thousands of HADS scores collected by the Symptom Monitoring Service.

In some cases these data consist of several repeated scores from the same patients. We are interested in the distribution of a HADS score obtained at random from the population of outpatients where each patient is weighted equally. To obtain the distribution of HADS scores we will therefore sample one score randomly from each patient in the SMS database, then record the observed frequencies. The empirical distribution obtained in this way is presented in Table 4.3.

Table 4.3. Empirical distribution of HADS scores among 17,599 cancer outpatients seen by the Symptom Monitoring Service.

Outcome H_0	Probability $P(H=H_0)$	Cumulative probability $P(H\leq H_0)$	Outcome H_0	Probability $P(H=H_0)$	Cumulative probability $P(H\leq H_0)$
0	0.037	0.037	22	0.016	0.929
1	0.049	0.085	23	0.013	0.943
2	0.054	0.139	24	0.010	0.953
3	0.056	0.196	25	0.009	0.962
4	0.060	0.255	26	0.007	0.970
5	0.065	0.320	27	0.006	0.976
6	0.059	0.379	28	0.005	0.981
7	0.059	0.438	29	0.004	0.985
8	0.054	0.492	30	0.003	0.988
9	0.051	0.544	31	0.003	0.991
10	0.049	0.592	32	0.002	0.993
11	0.043	0.636	33	0.002	0.995
12	0.041	0.676	34	0.002	0.996
13	0.037	0.714	35	0.001	0.997
14	0.035	0.749	36	0.001	0.998
15	0.034	0.783	37	0.001	0.999
16	0.029	0.812	38	0.000	1.000
17	0.024	0.836	39	0.000	1.000
18	0.021	0.857	40	0.000	1.000
19	0.021	0.878	41	0.000	1.000
20	0.019	0.897	42	0.000	1.000
21	0.017	0.913			

Having obtained the empirical distribution of HADS scores we can proceed to evaluate $g(h_c)$ and $G(h_c)$. These are the theoretical density and cumulative distribution functions of a truly continuous score. Of course with a truly continuous random variable, H, the probability of H taking the exact value h_c is zero. For a continuous H therefore

$$E[H_1 - H_2 \mid H_1 > h_c] = E[H_1 - H_2 \mid H_1 \geq h_c]$$

This is clearly not the case when H is discrete, and the challenge is therefore how to choose h_c. We wish to determine $E[H_1 - H_2 \mid H_1 \geq 15]$, which for a discrete H is equivalent to $E[H_1 - H_2 \mid H_1 > 14]$. But letting $h_c = 14$ means evaluating $g(h_c)$ and $G(h_c)$ at $h_c=14$.

For the discrete case, $G(14) = p(H \leq 14)$. Therefore $1\text{-}G(14)$ is the proportion of patients who score *15* or more, which is the quantity we are seeking. On the other hand $g(14)$ is the probability of a patient scoring *14* on the HADS, whereas we wish to evaluate g exactly where the cut-point is.

Suppose that underlying each observed HADS score, H, is a true continuous score, Z, and that $H=H_0$ when $H_0 - 0.5 < Z < H_0 + 0.5$. Patients would then score above the cut-off when $Z > 14.5$, and below otherwise. The values of g evaluated at $h_c=14$ and $h_c=15$ are *0.03540* and *0.03415* respectively. In approximating the continuous curve at $h_c=14.5$ we shall let $g(14.5)$ assume the middle value of *0.03478*. From table 4 we also find that $G(14) = 0.749$ (or equivalently that $G(14.5) = 0.749$ under the above scenario for an underlying continuous mechanism).

The expected change caused by regression to the mean can then be estimated as

$$E[H_1 - H_2 \mid H_1 > h_c] = (1-\rho)\,\sigma_t^2 \,\frac{g(h_c)}{1-G(h_c)}$$

$$\Rightarrow \qquad E[H_1 - H_2 \mid H_1 > 14.5] = (8.754)\,\frac{0.03478}{0.25104} = 1.213$$

That is, we would expect patients to score lower on the follow up assessment by an average of *1.21* units on the HADS scale. Any reduction of this magnitude may therefore be attributed to regression to the mean rather than an actual difference in HADS scores on the two occasions.

4.9 Quantifying uncertainty

The estimate of the RTM effect obtained above is subject to random variation. The estimated effect size is a function of four covariance parameter estimates as well as the two estimated functions g and G evaluated at h_c. It was therefore important to assess the accuracy of the result.

To do so, Model 3 from section 4.7 was applied with 500 bootstrap samples to obtain 500 sets of covariance parameter estimates. Similarly, the distribution of HADS scores was derived repeatedly from 500 bootstrap samples to obtain 500 estimates of g and G evaluated at h_c. Together these bootstrap estimates were used to obtain 500 estimates of the RTM effect which, when arranged in ascending order, could be used to obtain a 95% quantile-based confidence interval for the true RTM effect.

Using this method we found that the central 95% quantile-based confidence interval ranged from *1.02* to *1.43*.

4.10 The subscale dimensions

We had hypothesised that any additional distress experienced by patients in clinic, either as a result of the clinical environment or due to concerns about the possible outcomes of their upcoming appointment, would cause an increase in the anxiety scores but have less of an influence on the depression scores.

We found that the anxiety subscale scores fell by 1.26 points on average compared with just 0.48 points on the depression subscale. But does the larger drop on the anxiety subscale reflect fundamental differences between these two dimensions? Alternatively, could it simply be that the regression-to-mean effect is larger on the anxiety subscale?

In addressing this question one would need to estimate the expected change in the subscale scores conditional on the HADS total score from the first assessment being equal to or higher than 15. If we let A_1, D_1, A_2 and D_2 denote the anxiety and depression subscale scores on two occasions respectively and further let $H_1 = A_1 + D_1$ be the HADS total score as before, we would then need to determine

$$E[A_1 - A_2 \mid H_1 \geq 15]$$

to estimate the RTM effect on the anxiety subscale, and

$$E[D_1 - D_2 \mid H_1 \geq 15]$$

to estimate the effect on the depression subscale. Clearly, the estimation task is complicated by the higher dimensionality of the problem. Not only does the size of the effect depend on the joint distribution of longitudinal scores over time; it also depends on the joint distribution of the subscale scores. The problem of estimating the RTM effect individually for each subscale will not be pursued any further here.

4.11 Discussion

The audit described in this chapter was conducted to evaluate the strategy of asking patients to fill out a distress questionnaire while waiting for their appointment in clinic. The screening service routinely contacted patients who scored high in clinic for further assessment; it was therefore important to find out if patients' HADS scores were artificially inflated due to transient anxiety in anticipation of the upcoming appointment since identifying a large number of false positives would be inefficient.

When patients who scored 15 or more on the HADS in clinic were asked to complete the scale at home, a week later the mean change in scores was a reduction of 1.74 units (95% CI from 1.09 to 2.39) from the clinic visit. 72.5% (158/218) of the patients remained high scorers. However it was estimated that patients would score 1.21 units lower (95% CI from 1.02 to 1.43) on reassessment because of regression to the mean. This means that most of the observed drop could be accounted for by regression to the mean rather than an actual difference in scores on the two occasions.

In conclusion, there was evidence of a good agreement between scores obtained in clinic and scores obtained from patients at home one week later. The method of asking patients to complete a distress questionnaire while they wait for their appointment in clinic appears to be a reasonably reliable one.

4.11.1 Limitations

There were limitations to this study. We analysed data collected by a depression screening service operating in cancer clinics; the findings may not therefore generalise to other settings. The service only administered a second HADS to patients who had scored high in clinic. However, by analysing the data appropriately, taking into account the regression to the mean, the aims of the study were addressed adequately using the available data. The service only administered the second HADS as part of their follow-up call for a one month period; some patients could not be contacted during this limited period. However, the characteristics of patients on whom we had analysable data and those on whom we did not were similar; systematic bias is therefore unlikely. There may be limits to the intrinsic test-retest reliability of the HADS (as opposed to real changes in symptoms) but this is unlikely to be large over this time period or to represent a systematic bias. Finally, we were not able to assess the content of patients' clinical consultations, the nature of which might have accounted for some of the changes in scores; for example, if patients had been given good news this may have contributed to a fall in the HADS score and if they were given bad news, to a rise. Although the average change in scores was small the intra-patient variability was high with some patients scoring very differently on reassessment. It is possible therefore that a minority of patients are affected considerably by the clinic surroundings while the majority of patients remain unaffected. At the individual patient level we cannot rule out the possibility of an important 'clinic effect' in some cases.

The robustness of the RTM effect estimate was assessed in terms of sensitivity to random error. However sensitivity to a different model for the covariance was not assessed. Alternative models would yield different estimates for the covariance parameters. Nonetheless, the model selected fitted the data significantly better than the more parsimonious models, and the estimates from the model used were in good agreement with observed correlations in the SMS data. Even if the true correlation between repeated measurements obtained one week apart was as high as 0.90, the RTM effect would still be around 1.12. Similarly, a true correlation coefficient as low as 0.75 would still only result in a RTM effect of around 1.34.

We applied the Das & Mulder approach for continuous random variables to the quantitative, but discrete HADS scores. This was a particular problem when deciding how to choose h_c, the cut-off used to indicate a high score. We considered the nature of a theoretical underlying continuous HADS score and argued that a dichotomisation of the continuous process, that separates subjects into those who score below and those who score equal to or above 15 on the discrete scale, should be somewhere between a score of 14 and 15. We used the Das and Mulder approach with some simulated data from a continuous random variable that had been discretised in the way postulated in section 4.8 for the theoretical continuous HADS score. The method produced very similar results whether applied with the discrete or the continuous data.

We also assumed that the measurement errors or intra-patient variations were normally distributed. While this assumption is unlikely to hold entirely, it is at least likely that the intra-patient variations are symmetrically distributed about their means when S, the patient's true score, is close to 15.

The fitted model assumed homoscedastic error terms. We did not investigate this assumption although it is quite possible that individuals with elevated levels of distress were subject to greater intra-patient residual error. Since the variance parameters were modelled on the entire sample, the error variance and consequently the RTM effect are possibly somewhat underestimated as a result. However the presence of such a bias would not contradict our finding that the RTM effect accounted for most of the observed change in scores.

Finally, it is possible that the method of administration of the HADS instrument produces an effect. Perhaps patients scored differently on follow up because the HADS was read out to them over the telephone? As mentioned in section 4.3 this effect is likely to be small. Moreover, from the point of view of aiming to provide an efficient symptom screening service it is almost immaterial what factors caused the patient to score lower on follow-up. What is important is that the method of handing

out a questionnaire in clinic is a reasonably reliable one for identifying patients with symptoms of distress.

4.11.2 In context

The findings from the SMS audit are of relevance to the present project because they provide evidence that HADS scores obtained from patients in clinic are comparable with scores obtained from patients at home over the telephone, and may be used to identify patients likely to suffer from long-term distress. The problem of identifying patients likely to suffer from persistent distress is at the centre of the POD Study which is the topic of Chapter 5.

5 ANALYSIS OF THE POD STUDY

The POD Study addresses some very concrete examples of research questions that could potentially be addressed using the SMS data. Because PODS was specifically designed to address these questions the study may be regarded as providing the correct answers in the sense that any analysis addressing these questions using SMS data should lead to similar findings.

5.1 Background

Symptoms of psychological distress are common in cancer patients. Some patients with such symptoms develop major depression and are likely to require treatment. Less is known about the development of symptoms in patients who are identified with psychological distress, but who are not clinically depressed. Are their symptoms likely to persist and therefore to require treatment? And do most of these patients eventually develop major depression?

The POD Study was designed to investigate the course of symptoms over time in cancer outpatients who had been identified with distress by the SMS in a screened oncology outpatient clinic during a medical appointment, but who did not meet the criteria for major depression. A main aim of the study was to examine the persistence of patients' distress symptoms over an extended period. The study also sought to determine the demographic, disease and early distress characteristics that might predict cases of persistent distress. In particular it was thought that a second 'high reading' would be instructive, a confirmation of the patient's elevated distress level some time after the initial assessment. Another question of particular interest was therefore when to re-screen. What is the optimal time lag between the initial assessment and the confirmatory reading?

5.2 Design overview

Patients were followed up for a period of around seven months after the initial oncology outpatient clinic appointment. Patients who had scored 15 or more on the self-reported Hospital Anxiety and Depression Scale (HADS) in clinic (and who were subsequently found not to have major depression) were contacted by telephone

approximately four weeks later and asked to participate in the study. Patients who gave their consent proceeded to complete the first follow-up questionnaire and were subsequently asked to complete the questionnaire again over the telephone at approximately eight weeks, 16 weeks and 28 weeks after the initial clinic appointment.

5.3 Procedures

5.3.1 Screening

Patients were recruited into the study between June 2009 and April 2010. Patients were identified through the Symptom Monitoring Service (SMS) which operated in a large number of cancer outpatient clinics in NHS Lothian and NHS Greater Glasgow and Clyde in Scotland. The procedures of the SMS were described in detail in section 4.1.

Patients who scored 15 or more on the HADS in clinic were telephoned one or two weeks later and asked to complete a clinical interview for depression using the SCID (Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders (DSM-IV); First et al., 1999). The SCID interviews were tape recorded for quality purposes and were conducted by trained nurses or psychology graduates under the supervision of a psychiatrist. A small number of patients who had been short listed for this interview either refused, were excluded on medical grounds, or because they had recently had an interview for depression, or could not be contacted within a reasonable time period.

Of the patients who were interviewed, approximately one third were found to meet the criteria for Major Depressive Disorder (MDD). When a patient was identified with MDD a letter was issued to the patient's General Practitioner informing them of this. In addition, if found eligible, the patient was also offered trial participation in one of the SMaRT-Oncology trials of specially designed interventions for patients with depression and cancer (SMaRT-Oncology 2, ISRCTN: 40568538; SMaRT-Oncology 3, ISRCTN: 75905964). Descriptions of the two trials can be found in the published protocols (Walker, Cassidy & Sharpe, 2009a, 2009b).

The contact details of those who did not meet the criteria for MDD were passed on to the POD Study team with the permission of the patients. To be eligible for participation patients had to be aged 18 years or more and have a predicted survival, estimated by their cancer specialist, of 12 months or more. Eligible patients were sent a patient information leaflet informing them of the POD Study and were contacted within a couple of weeks by the PODS research team and asked to participate in the study.

5.3.2 Data collected at follow-up

The questionnaire completed in clinic was used as the baseline measure. The follow-up questionnaires at four, eight, 16 and 28 weeks contained the same scales as the SMS questionnaire. In addition to the questions about pain, fatigue and disturbed sleep the follow-up questionnaires also asked about dry mouth, lack of appetite, numbness or tingling, shortness of breath and drowsiness. The follow-up questionnaires also asked about nausea and vomiting separately. In addition to the questionnaire data the PODS team obtained information from patients' medical records after consent had been obtained at four weeks. Data on significant clinical events that occurred during the follow-up period were also obtained at the end of the study.

5.3.3 Sample size

A total of 325 patients were enrolled in the study during the recruitment period. The recruitment target of 350 patients was based on the number needed to obtain adequate stability in a logistic regression analysis at 28 weeks involving 10 predictors. It was assumed that around one third of patients would still be distressed on the final assessment. With 350 patients enrolled, and an expected attrition rate of 15% over the follow-up period, we anticipated around 100 events (distress cases) at 28 weeks.

5.4 Development of the study aims

An important question was whether most patients identified with distress in clinic developed persistent distress. To address this question it was necessary to define the meaning of persistence. Patients' HADS scores were measured at the oncology outpatients clinic visit and again at four, eight, 16 and 28 weeks afterwards. One very stringent criterion for meeting persistence would require that a patient score high (15 or more on the HADS) on all four follow-up assessments. However this would exclude anybody who fell just below the threshold on any of the assessments. A less exclusive rule could define a patient with persistent distress as someone who had scored high on the HADS on at least two occasions, although this rule could potentially assert someone as persistently distressed who had not scored high since week eight. Other alternative definitions included having a high score on both of week 16 and week 28 and having an average score of 15 or more across week 16 and 28. We defined a patient with persistent distress as someone with a single high score at 28 weeks. This decision was based on a number of considerations. It was seen as a suitably uncomplicated endpoint that would allow for a clear clinical message in a published article. The definition had face validity since persistently distressed patients were required to have significant symptoms of distress at the end of the follow-up period. Having significant distress symptoms at clinic appointment and again seven months later was judged to be a likely indicator of serious clinical distress. Finally the number of missing data points were minimised using this definition since only a single observation at 28 weeks was required for a patient to be included in the analysis of persistence.

Originally the study also sought to determine the number of patients who had developed major depression at 28 weeks. A depression screening interview was therefore conducted with all patients who had scored high on the HADS at the final assessment. Unfortunately, due to limited resources we were unable to carry out these clinical interviews within a reasonable time frame. This study aim was consequently abandoned.

Another important aim of the study was to identify baseline demographic, clinical and early distress characteristics that might be associated with an increased probability of persistent distress. We hypothesised that repeating the screening one month after the oncology outpatient clinic would be useful in predicting long-term distress. Identifying such early predictors could potentially inform the recruitment strategy of a trial of an intervention for significant psychological distress.

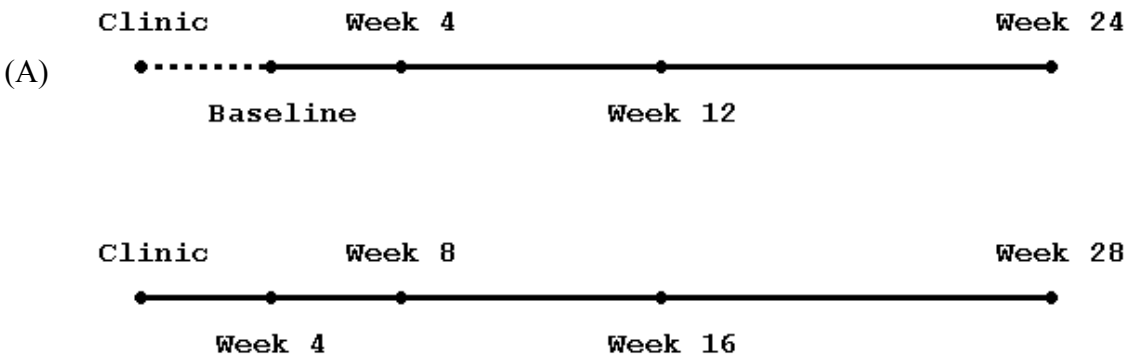
Finally, and somewhat unrelated to the present research project, the POD Study also sought to describe the quality of life and symptom burden of the study sample over the follow-up period. The questionnaires rated patients' quality of life using the EQ5D and the symptom burden using the CSQ (see below). The analysis of the EQ5D and CSQ outcomes is not of direct relevance to this project and will not be presented here.

5.5 Some design considerations

5.5.1 Two alternative ways of anchoring the follow-up times

Initially the intention was to only enrol patients who had scored high on the HADS on *two* separate occasions, first in the screened oncology outpatient clinic and second at a subsequent, confirmatory assessment carried out over the telephone approximately four weeks later. This was to avoid following up a large number of patients with transient distress whose symptoms were likely to stabilise within a short period of time once away from the hospital surroundings.

However, before the recruitment phase commenced we decided to follow-up all eligible patients who had scored high on the clinic assessment without the need for a second high score. Having to score high on the second assessment placed a restriction on the pool of eligible patients. It also became clear that we needed to know how symptoms evolve in patients with early signs of transient distress as it was unclear when would be the optimal time to re-screen. An illustration of the designs is given in Figure 5.1.



(A)

(B)

Figure 5.1. The PODS design with nominal time points anchored at (A) the first follow-up telephone call and (B) the oncology outpatient clinic visit.

The initial plan was to carry out all analyses with reference to the first telephone assessment as the baseline assessment (approximately four weeks after the clinic assessment). Patients' follow-up times were therefore calculated relative to the time of the four-week assessment rather than the time of the clinic visit. Figures 5.2 and 5.3 illustrate the spread of the actual follow-up times around the nominal time points when times are anchored at the clinic visit, and when anchored at the first follow-up telephone call respectively.

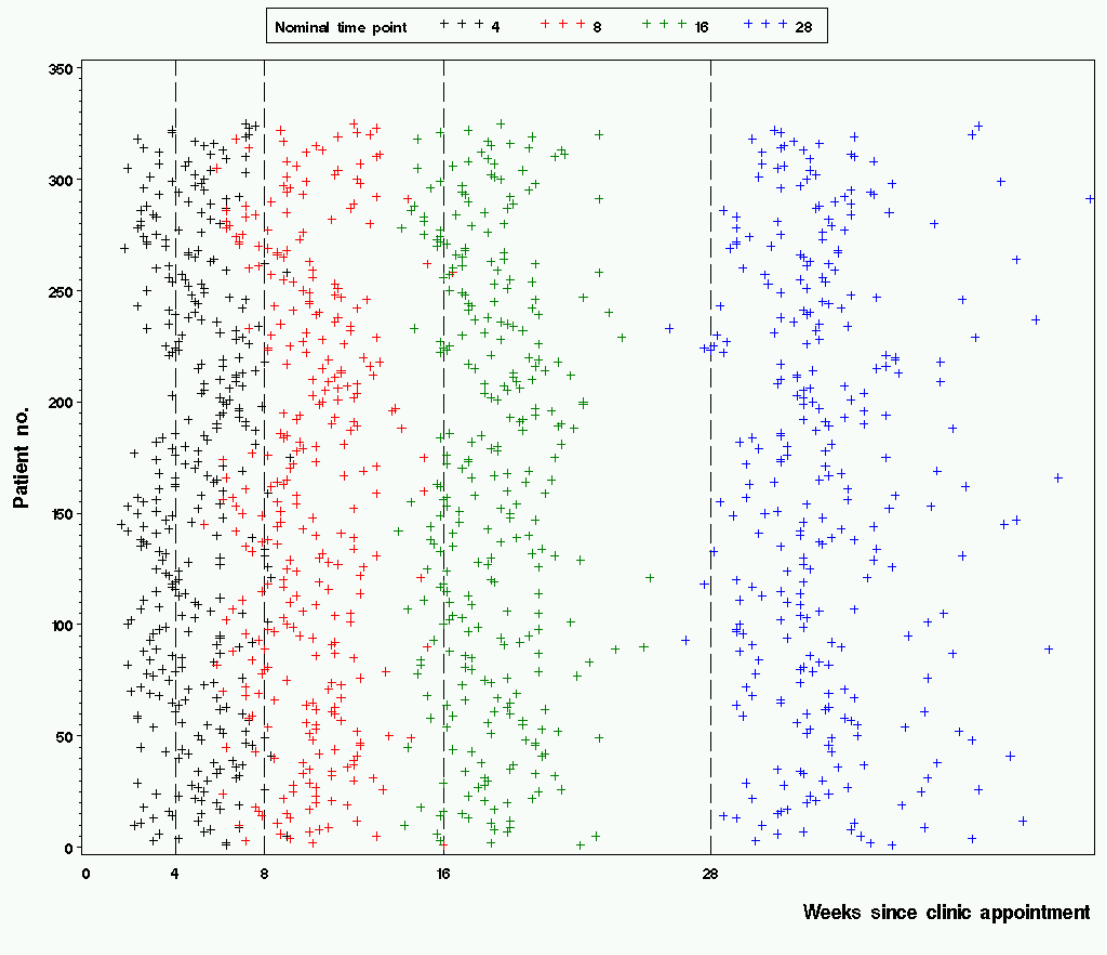


Figure 5.2. Actual follow-up times plotted for the entire sample when calculated from the time of the oncology outpatient clinic visit.

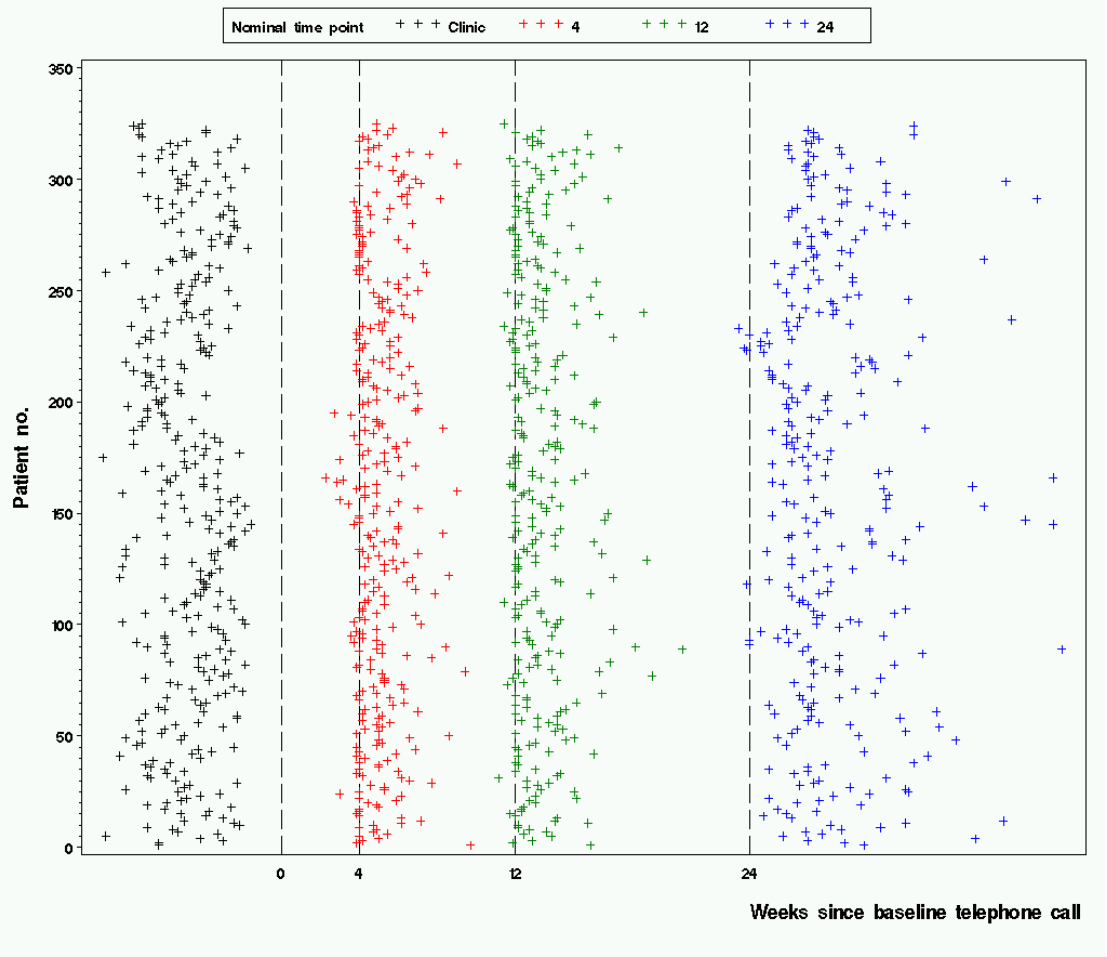


Figure 5.3. Actual follow-up times plotted for the entire sample when calculated from the time of the first follow-up telephone call.

It is clear that there is a considerably larger spread in the follow-up times when these are calculated from the time of the clinic visit. This is not surprising however. When follow-up times are anchored at the week-four visit, the variance of the actual follow-up times is simply *Var(X)*, where *X* is the time between the week-four assessment and the time of the follow-up assessment in question. In contrast when times are anchored at the time of the clinic visit the variance of the actual follow-up times is given by *Var(X+Y)*, where *Y* is the time between the week-four assessment and the clinic visit, and *X* is defined as before. Therefore, unless *Var(Y)=0* a post-hoc re-anchoring of the nominal time points will inevitably cause an increase in the variance of the actual follow-up times. In the present analysis all follow-up times are calculated with reference to the time of the clinic appointment.

5.5.2 The six week appointment

Initially a further follow-up assessment was planned six weeks after the clinic appointment. This assessment was quickly dropped from the design after recruitment had commenced because the spread of the actual follow-up times around the nominal time points was larger than anticipated thereby making it unfeasible to assess patients at all of four, six and eight weeks. Missing assessments from week eight have been replaced with assessments from week six where available. The small number of remaining outcomes from week six will not be used in the analysis.

5.6 External validity

The analysis of the POD Study presented in this chapter is focussed on the handling of missing data, i.e. the treatment of unobserved outcomes from patients who are already enrolled in the study. How we handle missing data can have important implications for the internal validity of a study. But internal validity is not the only thing needed for a study to lead to meaningful findings. Equally important is the question of where the study participants came from and what selection procedures took place before the arrival of the final sample ultimately enrolled and followed up.

The design was such that patients were only asked to participate if they were available to be contacted for the first follow-up telephone call four weeks after the clinic visit. As a result we have complete data on each of the first two occasions. However this particular presentation of the design is somewhat misleading as it masks the group of patients who, on the basis of their clinic visit, were eligible to participate, but who could not be contacted, or declined participation at four weeks. The studied sample therefore consisted of adult cancer patients, with a good prognosis, who scored high on the HADS in clinic without meeting the criteria for MDD, and who were available (and willing) to give consent four weeks later.

While this is no different from all other studies where only consenting patients are followed up, the one exception is the added limitation that eligible patients had to be available to be contacted via telephone within a narrow time window approximately four weeks after they had attended clinic. This added requirement made the inclusion

criteria rather ambitious and narrowed down the pool of eligible patients considerably. Figure 5.4 describes the flow of patients starting from the overall number screened by the Symptom Monitoring Service to the number of patients ultimately enrolled in the study.

5.7 Main analysis

5.7.1 Participants

A total of 567 patients were found eligible for study inclusion after completing the symptom screening questionnaire and the interview for depression. Of these, 138 (24%) declined the offer of participation, and 104 (18%) could not be contacted within the limited time window. Three hundred and twenty five patients were subsequently enrolled into the study. Basic demographic, cancer and distress characteristics collected during the oncology outpatient clinic visit by the SMS on all eligible participants and non-participants suggested that the study participants were younger by two years on average than patients who did not enrol. The two groups were remarkably similar on the other variables including HADS scores (Table 5.1).

Sixty one percent of the 325 enrolled patients were recruited from NHS Greater Glasgow and Clyde; the remaining patients were from NHS Lothian. The mean age was 61 (range 26 to 90) years. Eighty six percent of patients were female with the majority recruited from breast (57%) and gynae oncology clinics (23%). At baseline the majority of patients were disease-free (70%) with a median time since diagnosis of one and a half years. In the two months prior to screening 37% of patients had received no treatment, 34% had received chemo- or radiotherapy and 21% had received hormone treatment only. The remaining 9% had had surgery only (Tables 5.1 and 5.2).

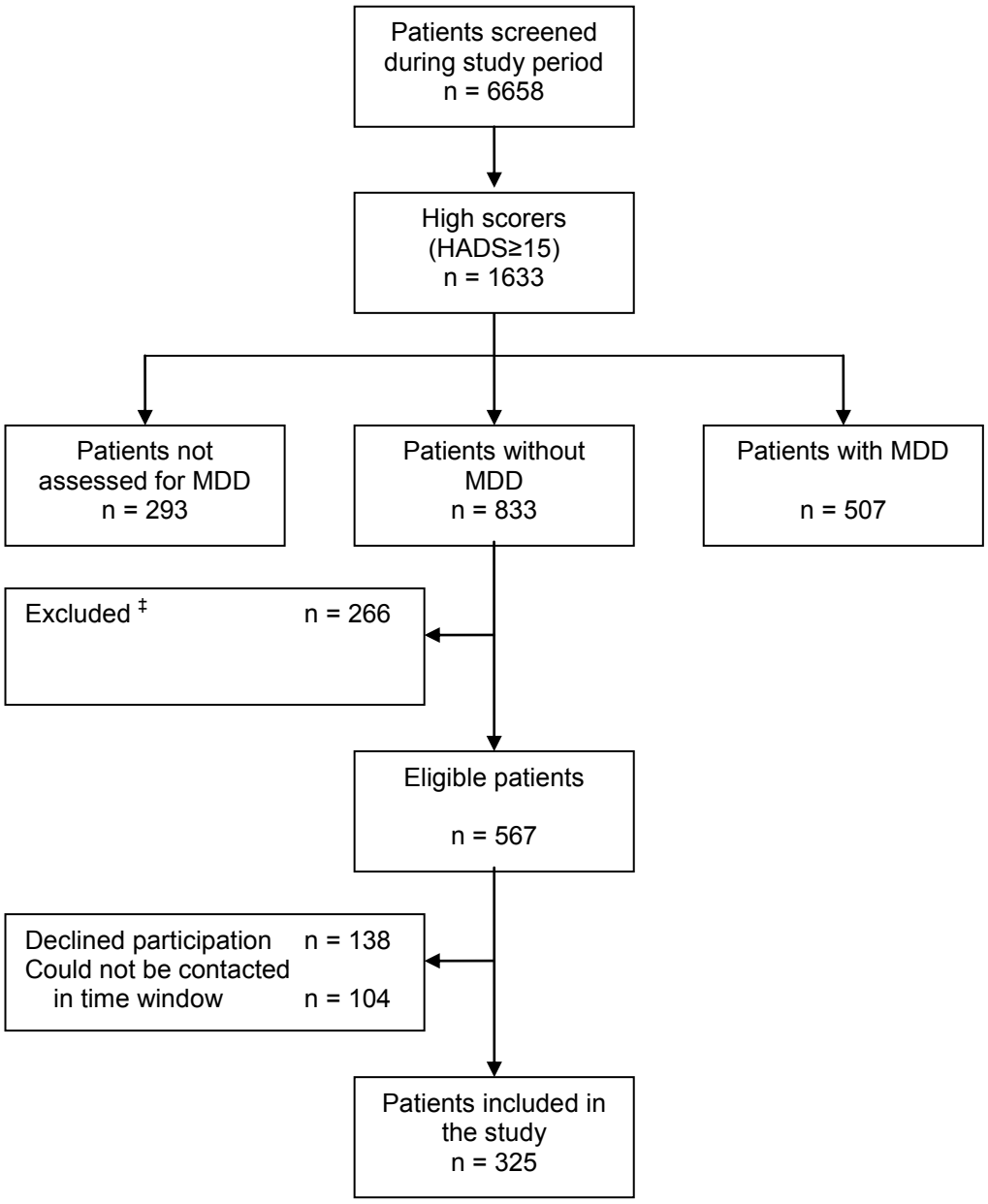


Figure 5.4. Recruitment flowchart.‡ Exclusion reasons were poor prognosis (<12 months) or prognosis unknown, patient too unwell to take part, patients referred outside study window and other reasons.

5.7.2 Mean follow-up times

Patients were scheduled to be followed up approximately four, eight, 16 and 28 weeks after the clinic visit. The distributions of the actual follow-up times were shifted somewhat away from the nominal assessment times (Figure 5.2). The four week assessments were completed on average 4.9 (IQR 3.6 to 6.1) weeks after the clinic visit. The eight, 16 and 28 week assessments were completed on average 10.0 (8.6 to 11.4) weeks, 18.2 (16.4 to 19.7) weeks and 33.1 (31.1 to 34.4) weeks after the clinic visit.

5.7.3 Marginal distribution of HADS scores at each time point

Figure 5.5 presents the marginal distribution of HADS scores at each of the five time points. At the clinic visit the distribution is highly skewed as it has been truncated to include only the high scores. At subsequent assessments the distributions quickly appear more symmetrical. For comparison the overall distribution of HADS scores among patients screened by the SMS is presented in Chapter 4 (Figure 4.5).

5.7.4 Distress scores over time

Individual HADS profiles from each of the participants are plotted in Figure 5.6 revealing a high degree of variability in the scores both between and within patients. The (smoothed) mean HADS profile over the follow-up period is presented in Figure 5.7. The mean distress level in the sample appears to be declining slightly over the duration of the follow-up period.

The observed mean scores at each of the nominal time points are reported in Table 5.3. Also presented are the proportions of patients that scored at or above the threshold of 15 at each time point. At the end of the study 37.0% (108/292; 95% CI: 31.4 to 42.5%) of the patients who were followed-up still had significant psychological distress.

Table 5.1. Characteristics recorded in clinic on the study participants compared with eligible patients in clinic who did not take part in the follow-up study.

	Participating patients n = 325	**Eligible non-participating patients** n=242	**P-value** [a]
Age (years)			0.046
Mean (SD)	61.0 (12.4)	63.1 (12.2)	
Median (range)	62.4 (26.3 to 89.7)	63.6 (33.7 to 88.4)	
Age group			
< 50 years	68 (21)	39 (16)	0.086
50 to 64 years	129 (40)	86 (36)	
> 65 years	128 (39)	117 (48)	
Gender			0.452
Female	280 (86)	203 (84)	
Male	45 (14)	39 (16)	
Cancer clinic [b]			0.752
Breast oncology	184 (57)	125 (52)	
Colorectal	18 (6)	15 (6)	
GI-clinic	9 (3)	7 (3)	
Gynae oncology	76 (23)	58 (24)	
Miscellaneous [c]	11 (3)	8 (3)	
Urology	27 (8)	29 (12)	
HADS score in clinic			0.830
Mean (SD)	18.7 (3.7)	18.6 (3.6)	
Median (range)	18 (15 to 34)	18 (15 to 35)	
Study centre			0.798
Edinburgh	127 (39)	92 (38)	
Glasgow	198 (61)	150 (62)	

Data are number (%) unless otherwise specified. [a] *Age in years* and *HADS score in clinic* were compared using t tests. All other p-values are from chi-square tests. [b] Out-patient clinic where the patient was screened at baseline. [c] Miscellaneous clinics included sarcoma (n=6), melanoma (n=5), lymphoma/myeloma (n=1) and other (n=7).

Table 5.2. Disease characteristics at baseline.

	n (%)
Total n	325 (100)
Primary cancer [a]	
Bowel	26 (8)
Breast	184 (57)
GU (urology)	28 (9)
Gynae	72 (22)
Other	15 (5)
Disease activity	
Disease-free	228 (70)
Local disease	36 (11)
Metastatic disease	60 (18)
Persistence of cancer	
Yes	54 (17)
No	271 (83)
Recurrence of cancer	
Yes	29 (9)
No	296 (91)
Cancer treatment in the two months prior to screening	
No treatment	119 (37)
Chemo/radiotherapy	110 (34)
Surgery only	29 (9)
Hormone treatment only	67 (21)
Time since diagnosis (years)	
Mean (SD)	3.0 (4.4)
Median (range)	1.5 (0.0 to 39.6)
Marital status	
Married/in a relationship	219 (67)
Divorced	28 (9)
Separated	8 (2)
Widowed	34 (10)
Single	36 (11)
Patient-reported duration of distress	
< 3 months	160 (49)
3 to 6 months	52 (16)
6 to 12 months	43 (13)
1 to 2 years	25 (8)
> 2 years	44 (14)
Employment status	
In full-time employment	51 (16)
In part-time employment	32 (10)
Retired	153 (47)
Unemployed (due to ill health)	61 (19)
Unemployed (other reasons)	28 (9)

Patient-reported duration of distress and disease activity were missing for one patient.Date of diagnosis was missing for four patients. [a] *GU (urology)* included bladder (n=1), kidney (n=1), prostate (n=22), testis (n=3) and ureters (n=1); *Gynae* included cervix (n=14), fallopian tube (n=1), ovary (n=38), uterus (n=13), vagina (n=3), vulva (n=3); *Other* included haematological cancer (n=4), malignant melanoma (n=3), sarcoma (n=3), cancer of the liver (n=1), pancreas (n=2), and unknown sites (n=2).

Table 5.3. Mean HADS scores and number of cases with significant distress (HADS≥15) during the follow-up period.

	Clinic appointment	1 month	2 months	4 months	7 months
All available data					
n	325	325	307	296	292
mean (SD)	18.7 (3.7)	13.4 (5.7)	13.7 (6.4)	13.4 (6.5)	12.5 (6.7)
cases (%)	325 (100%)	130 (40.0%)	133 (43.3%)	120 (40.5%)	108 (37.0%)
Complete cases [a]					
n	269	269	269	269	269
mean (SD)	18.6 (3.6)	13.4 (5.6)	13.4 (6.2)	13.3 (6.5)	12.2 (6.6)
cases (%)	269 (100%)	107 (39.8%)	112 (41.6%)	109 (40.5%)	96 (35.7%)

[a] The analysis of complete cases is of the 269 patients for whom data were collected at every follow-up occasion.

5.7.5 Distress trajectories

One of the aims of the study was to characterise patients' distress trajectories over time. With 325 individual patient profiles this posed a challenge. To simplify the problem the scores were dichotomised at each time point again using the threshold of 15. Consequently, with four time points and two distress states at each time point there were a total of 16 ways in which a patient could progress over the period. (In fact there were three possible outcomes at each time point since scores might also be missing. However the analysis was restricted to complete cases only so as not to complicate the findings unnecessarily.)

The number and proportions of patients falling into each of the possible categories are shown in Table 5.4. Of the 16 possible trajectories two in particular were clearly very common: Not being distressed at any of the time points, and being distressed at all time points.

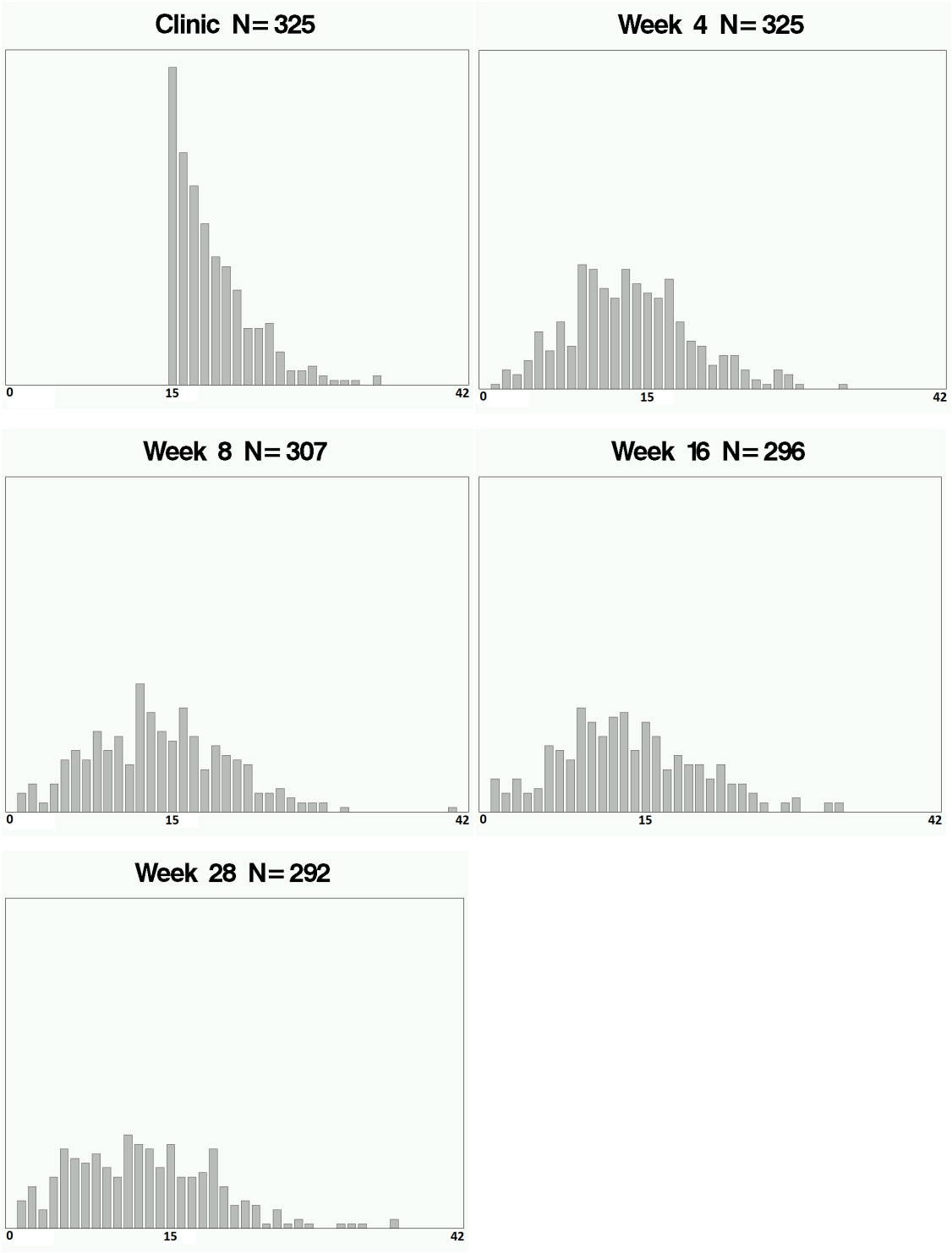


Figure 5.5. Marginal distribution of HADS scores in the PODS sample at each of the five time points.

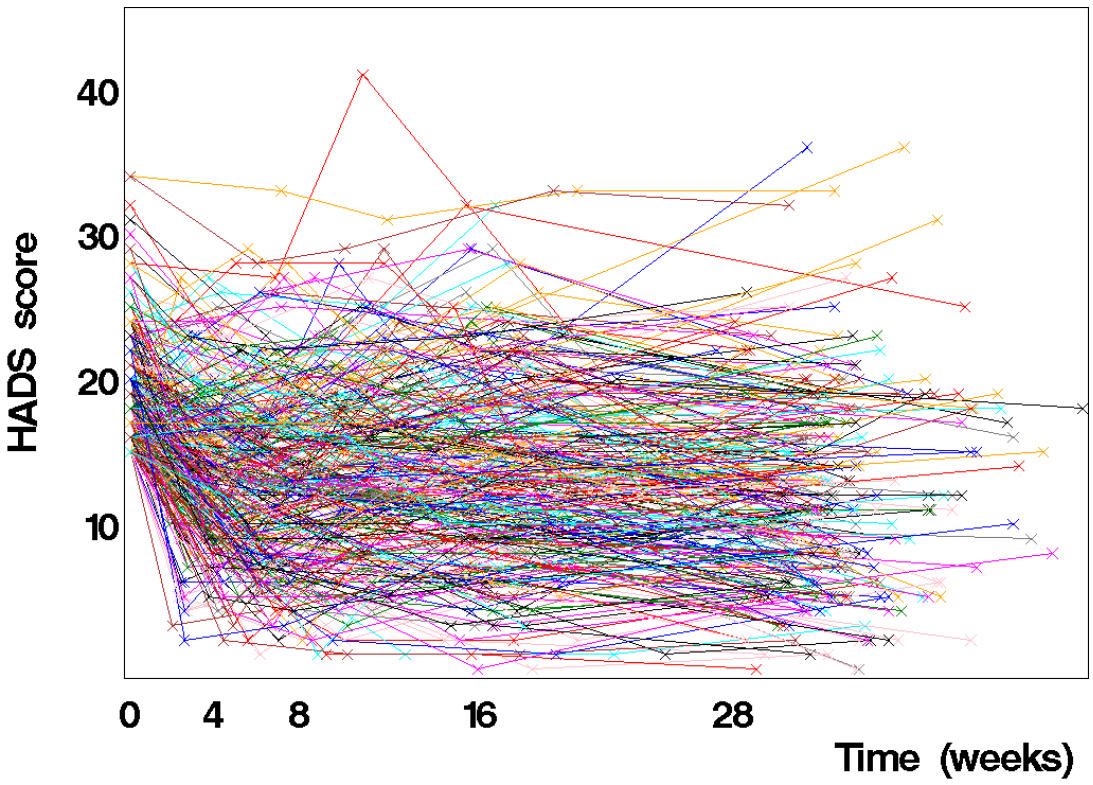


Figure 5.6. HADS profiles over the study period plotted in a single plot.

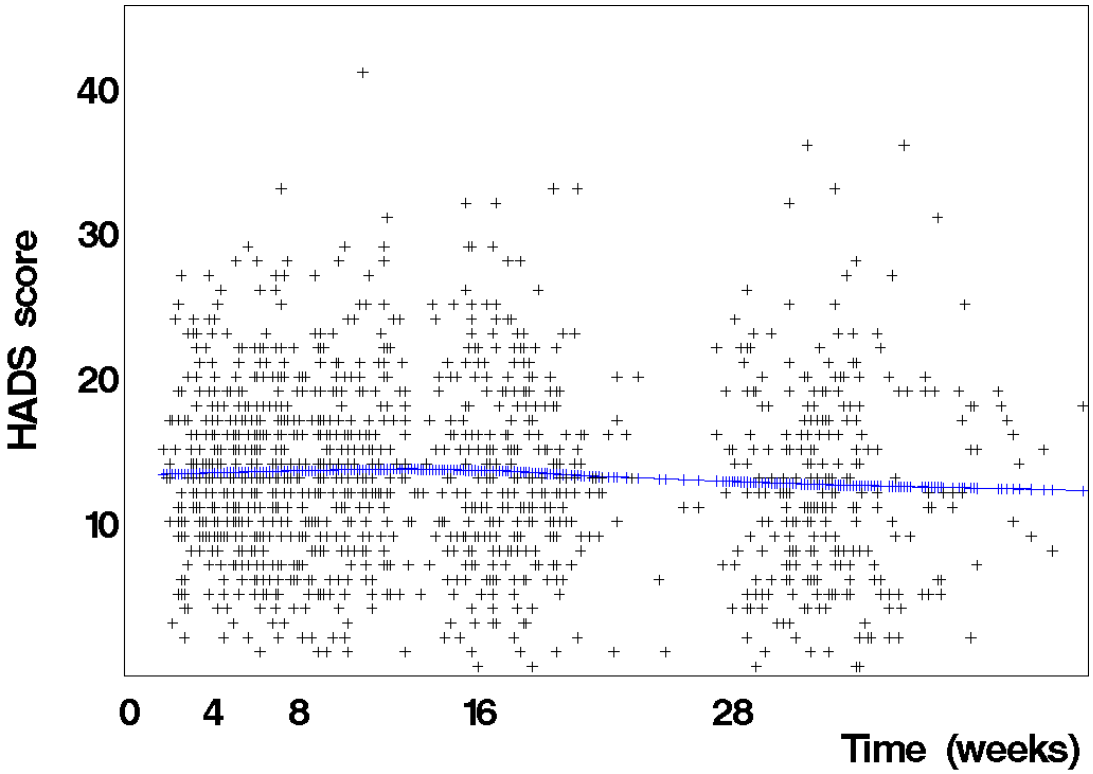


Figure 5.7. Smoothed (lowess) mean HADS profile over the follow-up period (continuous curve) and individual follow-up HADS scores (crosses).

Table 5.4. Variations in patient distress status over the four follow-up time points (completers only, n=269).

	No distress at follow-up	Distress at 8 weeks only	Distress at 16 weeks only	Distress at 28 weeks only	Distress at 8 & 16 weeks	Distress at 8 & 28 weeks	Distress at 16 & 28 weeks	Distress at 8, 16 & 28 weeks
Distress at 4 weeks								
Yes	12 (4)	15 (6)	2 (1)	4 (1)	14 (5)	2 (1)	6 (2)	52 (19)
No	104 (39)	5 (2)	13 (5)	11 (4)	8 (3)	7 (3)	5 (2)	9 (3)

Data are number (percentage)

Table 5.5. Distress persistence over time.

	Patients with complete data n=269
Distress on all four follow-up occasions	52 (19%)
Distress on three follow-up occasions	31 (12%)
Distress on two follow-up occasions	41 (15%)
Distress on one follow-up occasion	41 (15%)
Not distressed at any of the follow-up occasions	104 (39%)

5.7.6 Persistence of distress

The 269 patients with complete data were then grouped into just five categories according to the number of times they scored high on the HADS (Table 5.5).

The prevalence of distress cases in a similar general outpatient cancer population has been estimated to be around 22% (Sharma et al., 2011). But is it the same group of patients who remain distressed over time? We found that approximately 40% of patients score high on the HADS up to seven months after being identified with distress in clinic. It appears therefore that symptom levels do not simply revert to background levels on follow-up, but tend to persist in a particular group of patients. The data in Table 5.5 provide further evidence that the distress is recurrent in a consistent group of patients and that there is a large group of patients (19% of the sample) who are constantly subjected to elevated levels of distress.

Fifty eight percent of patients (156/269) fell into either of the two extreme categories of not being distressed at all or of being distressed at every time point. Under the hypothesis of statistical independence between distress status at four, eight, 16 and 28 weeks, conditional on the observed prevalences, we would expect far fewer patients in either of these two extreme categories (Table 5.6). The probability of observing frequencies as extreme as or more extreme than those actually observed is practically zero. This is hardly a surprising finding since the null hypothesis of independence between the repeated measurements is clearly an unreasonable one which illustrates the challenging problem of defining persistence.

Table 5.6. Distress persistence over time: Expected frequencies under the null hypothesis of statistical independence.

	Observed frequencies†	Expected frequencies	p-value
Distress on all four follow-up occasions	52	7.0	<0.0001
Distress on three follow-up occasions	31	41.8	
Distress on two follow-up occasions	41	93.4	
Distress on one follow-up occasion	41	92.5	
Not distressed at any follow-up occasions	104	34.3	

† patients with complete data only (n=269).

5.7.7 The utility of a confirmatory reading

It was originally hypothesised that a reassessment approximately four weeks after the initial high reading in clinic would be helpful in identifying patients with persistent distress. Of patients who scored high on the confirmatory reading at four weeks, 60% (71/118) remained distressed at 28 weeks, up from 37% in the overall sample of patients who scored high in clinic (Table 5.7). The sensitivity estimate of around 66% implies that a third of patients with distress at 28 weeks would be missed (i.e. would score below the threshold) in the confirmatory reading at four weeks. While the sensitivity can be somewhat improved upon by delaying the confirmatory reading until eight weeks, little is gained in terms of the positive predictive value. For comparison the diagnostic qualities of the outcomes at four and eight weeks together, and at 16 weeks are also presented, although waiting 16 weeks for a confirmatory reading is not a practical option.

Table 5.7. Distress status at four, eight, and 16 weeks as a diagnostic tool to predict persistent distress at 28 weeks.

	Sensitivity	Positive Predictive Value
In clinic	-	37.0
4 weeks	65.7	60.2
8 weeks	73.1	62.8
16 weeks	75.8	67.0
4 & 8 weeks	55.8	65.2

5.7.8 Associations with distress at seven months

Univariate analyses of associations with patient characteristics at baseline showed that patients who scored 20 or more on the HADS in clinic were more likely to be distressed at 28 weeks (Table 5.8). This was confirmed in the multivariable analysis with odds of persistent distress estimated to be 1.85 times higher in the group of patients who scored 20 or more in clinic (95% CI: 1.08 to 3.15). The type of cancer treatment in the two months prior to the clinic appointment was also found to be independently associated with persistent distress. Patients who had received chemo- or radiotherapy (OR: 0.47, 95% CI: 0.25 to 0.89) or who had undergone surgery (OR: 0.31, 95% CI: 0.11 to 0.87) were less likely to have persistent distress than patients who were not in receipt of treatment (p=0.034).

Overall the model did not reveal any particularly strong predictors at baseline and did not discriminate well between patients with and without persistent distress (c-index: 0.65). The overall likelihood ratio test of the full model compared to the null model did not reach statistical significance (χ^2=19.4, df=13, p=0.111) and individual significant associations should therefore be interpreted with caution.

Distress status at four weeks after clinic screening was subsequently added to the multivariable model in a separate analysis (Table 5.8; analysis (2)) and was shown to be a strong independent predictor of persistent distress (OR: 5.48, 95% CI: 3.13 to 9.60). Scoring 20 or more on the HADS in clinic carried little additional information having adjusted for distress at four weeks and was no longer a significant predictor.

We also sought to investigate how the occurrence of a significant worsening in clinical status over the study period might affect patients' distress levels. We defined such occurrences as either a worsening in disease status from disease-free to local disease, or from local disease to metastatic disease, or as a recurrence of cancer during the follow-up period. We found that there were too few such events to conduct a meaningful analysis, perhaps unsurprisingly, since the study sample consisted largely of patients on long-term follow-up.

Table 5.8. Prevalence of significant distress (HADS≥15) at seven months and associations with patient characteristics *(continued on next page)*.

Variable	Total *n*	Significant distress at 7 months *n (%)*	No/mild distress at 7 months *n (%)*	Univariate analysis Odds ratio *(95% CI)*	Univariate analysis P-value
Total	292	108 (37)	184 (63)		
Gender					**0.138**
Female	252	89 (35)	163 (65)	1	
Male	40	19 (48)	21 (53)	1.66 (0.85, 3.25)	
Age					**0.623**
< 50	60	25 (42)	35 (58)	1	
50 to 64	118	44 (37)	74 (63)	0.83 (0.44, 1.57)	
≥ 65	114	39 (34)	75 (66)	0.73 (0.38, 1.38)	
Primary cancer					**0.379**
Bowel	20	8 (40)	12 (60)	1	
Breast	173	65 (38)	108 (62)	0.90 (0.35, 2.33)	
GU (urology)	25	13 (52)	12 (48)	1.63 (0.49, 5.34)	
Gynae	60	18 (30)	42 (70)	0.64 (0.23, 1.84)	
Other	14	4 (29)	10 (71)	0.60 (0.14, 2.60)	
Disease activity [†] [ŧ]					**0.800**
Disease-free	218	80 (37)	138 (63)	1	
Active disease	73	28 (38)	45 (62)	1.07 (0.62, 1.85)	
Cancer treatment [¥]					**0.059**
No treatment	105	46 (44)	59 (56)	1	
Chemo/radiotherapy	95	29 (31)	66 (69)	0.56 (0.32, 1.01)	
Surgery only	28	6 (21)	22 (79)	0.35 (0.13, 0.93)	
Hormone treatment only	64	27 (42)	37 (58)	0.94 (0.50, 1.75)	
Marital status					**0.929**
Not married	91	34 (37)	57 (63)	1	
Married	201	74 (37)	127 (63)	0.98 (0.59, 1.63)	
HADS score at screening					**0.028**
Score < 20	196	64 (33)	132 (67)	1	
Score ≥ 20	96	44 (46)	52 (54)	1.75 (1.06, 2.88)	
Distress status 1 month after screening					**<0.001**
No significant distress	174	37 (21)	137 (79)	1	
Significant distress	118	71 (60)	47 (40)	5.59 (3.33, 9.38)	

Table 5.8. *(Continued)*.

Variable	Multivariate analysis [†] (analysis 1) [‡] Odds ratio *(95% CI)*	P-value	Multivariate analysis [†] (analysis 2) [‡] Odds ratio *(95% CI)*	P-value
Total				
Gender		**0.432**		**0.488**
Female	1		1	
Male	1.83 (0.40, 8.34)		1.78 (0.35, 9.09)	
Age		**0.383**		**0.635**
< 50	1		1	
50 to 64	0.85 (0.44, 1.65)		0.98 (0.48, 2.00)	
≥ 65	0.63 (0.32, 1.26)		0.75 (0.35, 1.58	
Primary cancer		**0.619**		**0.834**
Bowel	1		1	
Breast	1.25 (0.34, 4.61)		0.87 (0.21, 3.63)	
GU (urology)	1.16 (0.28, 4.77)		0.68 (0.15, 3.18)	
Gynae	0.87 (0.22, 3.48)		0.79 (0.18, 3.56)	
Other	0.51 (0.11, 2.41)		0.39 (0.07, 2.13)	
Disease activity [† ₣]		**0.779**		**0.410**
Disease-free	1		1	
Active disease	1.09 (0.60, 1.97)		1.31 (0.69, 2.49)	
Cancer treatment [¥]		**0.034**		**0.044**
No treatment	1		1	
Chemo/radiotherapy	0.47 (0.25, 0.89)		0.42 (0.22, 0.84)	
Surgery only	0.31 (0.11, 0.87)		0.38 (0.12, 1.14)	
Hormone treatment only	0.83 (0.42, 1.64)		0.86 (0.41, 1.81)	
Marital status		**0.878**		**0.718**
Not married	1		1	
Married	1.04 (0.60, 1.81)		1.12 (0.62, 2.02)	
HADS score at screening		**0.025**		**0.347**
Score < 20	1		1	
Score ≥ 20	1.85 (1.08, 3.15)		1.32 (0.74, 2.37)	
Distress status 1 month after screening		-		**<0.001**
No significant distress			1	
Significant distress			5.48 (3.13, 9.60)	

[₣] Disease activity at screening. [†] Disease activity was unknown for one subject in the group without significant distress at 7 months. Total n = 291. [‡] (1) including baseline information only, (2) Including distress status at 1 month as well as baseline information. [¥] Cancer treatment in the 2 months prior to screening. Surgery could have included hormone therapy; Chemo/radiotherapy could have included surgery and/or hormone therapy.

5.8 Missing data

An unusual feature of the design was that study participants were asked for their consent at the start of the four week follow-up telephone call instead of at the time of the first assessment (in clinic). As was noted above, the study sample was therefore not defined until the first follow-up time point; there were consequently no missing data from either the clinic visit or at four weeks. As for the remaining time points, 94% (307/325) of patients completed the questionnaire at eight weeks, and 91% (296/325) and 90% (292/325) of patients completed the questionnaires at 16 and 28 weeks respectively.

The analysis presented in section 5.7 is based on only partially observed data. To see why this is important we will consider initially the estimated proportion of patients still distressed after 28 weeks. This was estimated to be 37.0% and was based on the 292 patients with observed distress status at 28 weeks. Implicit in this estimate therefore is the assumption that the 33 patients with missing data had similar distress levels at 28 weeks to those patients whose distress was measured. This means that one would expect around 12 of the 33 patients with missing data to be distressed at 28 weeks.

To see if this is a reasonable assumption it is helpful to investigate how the 33 patients scored at earlier time points compared to the 292 patients with observed data. The mean HADS scores (and percentage with a high score) in the group of 33 were 18.8 (100%), 13.7 (36%), 15.8 (48%) and 15.2 (44%) in clinic, at four, eight and 16 weeks respectively. In comparison, the mean scores (percentage with a high score) in the 292 patients with complete data at 28 weeks were 18.6 (100%), 13.4 (40%), 13.5 (43%) and 13.3 (40%) at the same time points. While it is difficult to conclude anything on the basis of these numbers it would appear that patients with missing data at 28 weeks might experience somewhat higher distress levels at earlier time points. Table 5.3 supports this idea. The estimates from the complete cases are consistently lower than those based on the available cases suggesting that patients who comply more with data collection procedures are somewhat less distressed on average that those who do not. It therefore seems more reasonable to assume that

more than 12 of the 33 missing scores at 28 weeks would have been high had they been observed. In the following we shall assess the sensitivity of the findings in 5.7 to different treatments of the missing data.

5.8.1 Missing data patterns

The missing data patterns are tabulated in Table 5.9. Two hundred and sixty nine (83%) patients provided complete data at all time points. Of the 56 patients with at least one missing data point, 39 patients had just a single missed assessment, while the remaining 17 patients missed two (n=10) or three (n=7) assessments. The most frequently observed missingness pattern was having a single missed assessment at the final time point (n=17).

Table 5.9. Patterns of missingness over the four follow-up time points.

Pattern no.	Week 4	Week 8	Week 16	Week 28	Number (%) of patients with the indicated pattern
1	O	O	O	O	269 (83)
2	O	O	O	M	17 (5)
3	O	O	M	O	13 (4)
4	O	M	O	O	9 (3)
5	O	O	M	M	8 (2)
6	O	M	M	M	7 (2)
7	O	M	O	M	1 (0)
8	O	M	M	O	1 (0)

Note: O (M) indicates that the data point was observed (missed). E.g. a patient with missingness pattern 2 provided data at each of four weeks, eight weeks and 16 weeks, but not at 28 weeks.

The mean HADS trajectories are plotted separately for each of the missingness patterns 2 to 6 (Figure 5.8). Also shown in each plot is the mean trajectory from the complete cases for comparison. Figure 5.9 presents the mean HADS trajectories plotted separately for the complete and the incomplete cases (mean scores pooled across missingness patterns 2 to 8). The observed mean scores are higher on average in the incomplete cases. It seems we cannot assume similar levels of distress in patients with and without missing data.

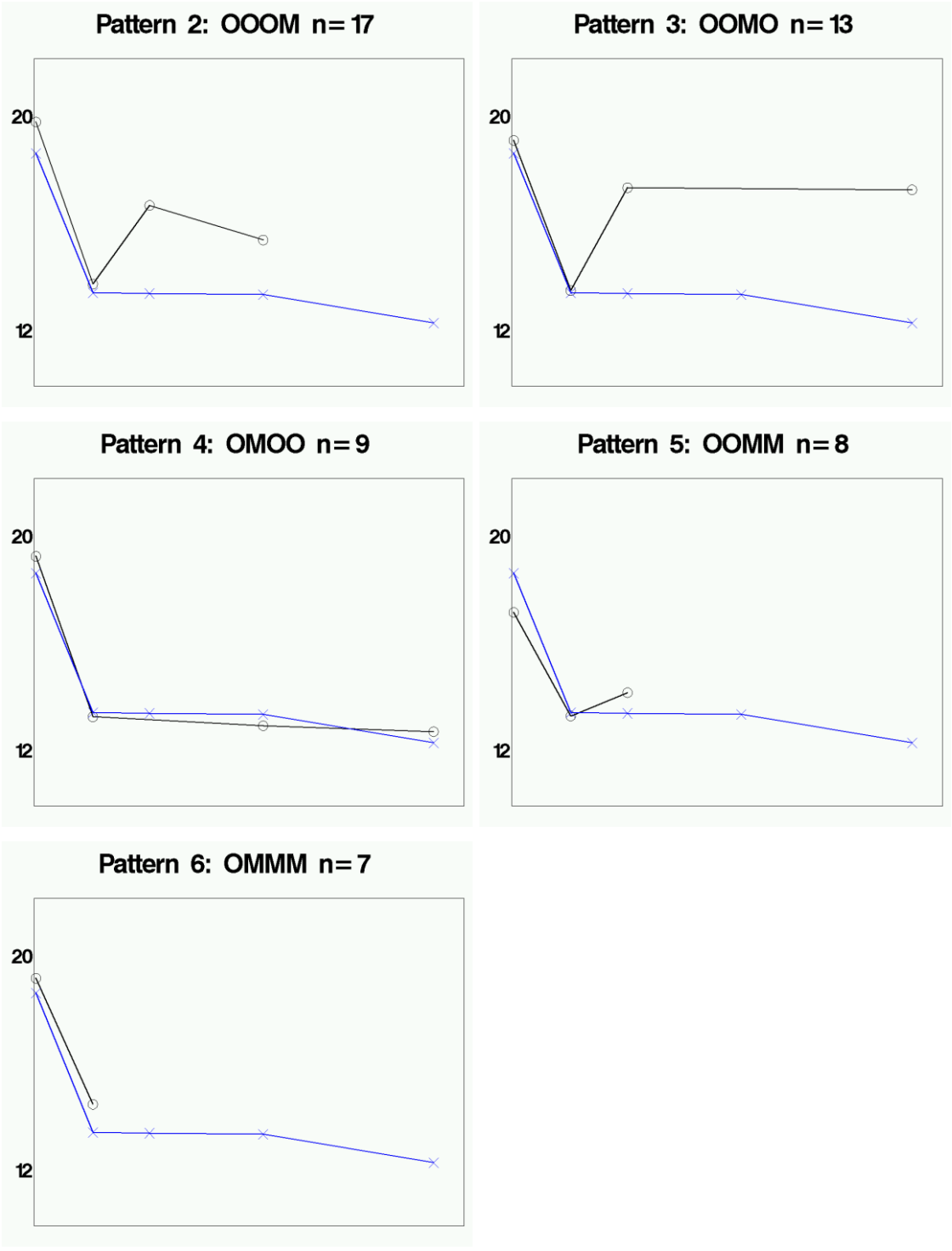


Figure 5.8. Mean HADS trajectories over the study period plotted separately for missingness patterns 2 to 6 (line segments connected by circles) alongside the trajectory from the complete cases (line segments connected by crosses). The four letter combinations denote the missingness patterns. E.g. OOMO indicates patients were observed at four, eight and 28 weeks, but not at 16 weeks.

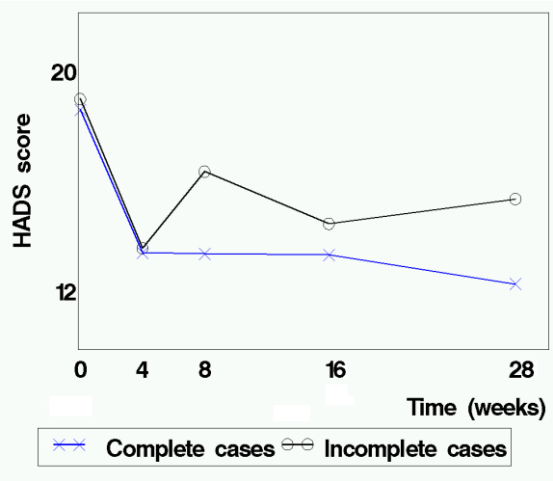


Figure 5.9. Mean HADS trajectories plotted separately for complete and incomplete cases. Sample sizes for the incomplete cases vary: week 0: n=56; week 4: n=56; week 8: n=38; week 16: n=27; week 28: n=23.

5.9 Reanalysis using Multiple Imputation

Patients with some missing data were likely to score higher on the HADS on those occasions when their scores were observed than patients with complete data on all occasions. As a result the MCAR assumption underlying the analysis presented above does not appear to be justified. A logical next step was therefore to conduct an analysis assuming the more general MAR mechanism for the missing data and assess the influence on the findings to see if the results presented were robust to a relaxation of the MCAR assumption.

The distributions of covariates were compared between patients with and without outcome data. All collected variables that were potentially relevant were analysed for associations with missingness. The following variables were included in the imputation model and used to predict the missing data: Gender; age in years (<50, 50 to 64, ≥65); primary cancer site (bowel, breast, GU (urology) gynae, other); disease activity (disease-free, active disease) and cancer treatment (no treatment, chemo/radiotherapy, surgery only, hormone treatment only) at study entry; marital status (married, not married); health board (NHS Lothian, NHS Greater Glasgow and Clyde); all available HADS scores collected previously or subsequently.

The missing HADS scores (range 0 to 42) were imputed 100 times each using the MI procedure of the SAS software (version 9.1). The imputations were generated using

Markov Chain Monte Carlo (MCMC) methods appropriate for non-monotone missingness patterns and were sampled from a single Markov chain with an initial burn-in sequence of 200 iterations, leaving 100 iterations between each imputation to avoid serial dependence.

The results from the multiple imputation analysis were broadly similar to those of the main analysis. The estimated mean HADS scores from the imputed datasets were 13.7, 13.6 and 12.7 at two, four, and seven months respectively. The levels of caseness were 43% at two months, 42% at four months and 38% at seven months. These estimates were somewhat higher than those based on the incomplete data suggesting that patients with higher levels of distress were more likely to have some observations missing. The analysis of the utility of a confirmatory assessment some time after the original screening episode in clinic also resulted in findings very similar to those from the main analysis. The results from the analysis of patient distress trajectories over time were also largely similar, although the estimated proportion of patients without significant distress at any of the follow-up occasions was reduced from 39% to 37% when taking account of the missing data using multiple imputation. Finally, results from the regression analysis were almost identical to those of the main analysis again confirming the finding that distress status at four weeks was the only strong predictor of persistent distress at 28 weeks.

In summary, the results from the multiple imputation analysis (under a MAR mechanism) were very similar to the results of the main analysis. The findings from the POD Study therefore appear reasonably robust despite the limited amount of missing data. In practice the MCAR assumption is probably rarely correct. Arguably therefore the MAR mechanism should be the working assumption with a robust sensitivity analysis subjecting the data to a range of alternative scenarios.

5.10 Discussion

We found that 37.0% of patients (95% CI: 31.4% to 42.5%) were still distressed at 28 weeks with little improvement in average HADS scores over the follow up period. The distress was recurrent in a consistent group of patients with around a fifth of

patients scoring high on the HADS on every occasion. We found that there were no strong predictors at baseline that would allow clinicians to discriminate well between patients at risk, and those not at risk, of becoming persistently distressed. However, a second high score approximately four weeks after the original screening episode increased the odds more than five fold of persistent distress half a year later. The introduction of a confirmatory reading should therefore be considered when screening for long-term distress.

There were some missing data, especially towards the latter part of the follow-up period when data completeness rates fell to around 90%. Exploratory analysis suggested that patients with missing data scored higher on the HADS on average than patients with complete data. A sensitivity analysis using multiple imputation to impute the missing values under a MAR mechanism was conducted. Analyses of the imputed datasets resulted in very similar findings to those of the main analysis. The findings therefore appear to be reasonably robust, certainly to a relaxation of the MCAR assumption.

5.10.1 Limitations

The response rate of eligible patients ultimately enrolled into the study was relatively low. As discussed in section 5.6 this was mainly due to the unusual design where only patients available to be contacted within a narrow time window could be enrolled. Comparisons of basic characteristics between eligible participants and non-participants showed that those enrolled were two years younger on average, but that there were no other differences on the variables compared. Indeed the HADS scores were remarkably similar between the two groups leaving little evidence on the basis of the comparisons that eligible non-participants would have progressed substantially differently from those who participated.

The study was carried out in cancer outpatients attending clinics in Edinburgh and Glasgow in Scotland, UK. The findings may not therefore generalise to other settings or medical conditions. As this was a longitudinal study of distress in cancer patients it would have been interesting to have investigated how significant clinical changes

in the patients' medical conditions during the follow up period affect distress levels. Unfortunately it was not possible to investigate this point since too few patients experienced such events to conduct a meaningful analysis. Finally, due to limited resources we were unable to ascertain the number of patients whose distress developed into major depression during the study.

5.10.2 Implications

A significant proportion of cancer outpatients who present at clinic with symptoms of psychological distress are still distressed half a year later. A second 'high reading' four weeks after the initial screening episode was found to be a strong predictor of long-term distress; the introduction of a confirmatory reading might help screening services identify patients likely to require treatment, and might also inform the recruitment of clinical trials in patients with psychological distress.

5.10.3 In context

The POD Study was a longitudinal study investigating the development of patients' distress scores over time. Patients were followed up at regular intervals, and retention rates were good with only 10% of patients failing to provide data after seven months. In contrast, the distress data routinely collected by the Symptom Monitoring Service are incomplete and unevenly spaced between patients, depending on the timing of their clinic appointments. Analysis with the screening data collected by the Symptom Monitoring Service is the topic of Chapter 7. But before that: some exploratory work carried out in preparation for such analysis.

6 BAYESIAN ANALYSIS WITH MISSING DATA USING WINBUGS

The present chapter does not set out to answer a research question, and the work reported is not intended to add value to the literature on analysis with missing data or Bayesian modelling using WinBUGS. Rather, the chapter is a narrative describing experimental work that was carried out by the author to gain experience necessary for the analyses presented in Chapters 7 and 8, and it represents an important step in the early development of the author's understanding of practical analysis with incomplete data. Some of the ideas presented in Chapter 3 are illustrated in the present chapter, and as such it may also be seen as an extension to Chapter 3.

The chapter is introduced with some exploratory work with alternative approaches to multivariate incomplete data modelling in WinBUGS. We study the POD Study data under a MAR assumption and lastly consider alternative non-response processes through analysis with simulated data.

6.1 Inducing dependency through a linear predictor

Assume that variable Y is a quantitative variable measured on two occasions, t_1 and t_2, resulting in two potential observations, Y_{i1} and Y_{i2}, from the ith subject. Further assume that Y_1 is fully observed and that Y_2 is missing in some cases. Y_1 and Y_2 are typically correlated since they are repeated measurements within the same subject. One way to express the dependency is through a linear regression of Y_2 on Y_1. Specifically we can model Y_2 using the following model:

$$Y_{i2} = \alpha + \beta Y_{i1} + e_i$$

Here α is the intercept, β is the regression coefficient, and the e_i are independent error terms with $e_i \sim N(0, \sigma^2)$.

6.1.1 Bayesian predictions

Under the Bayesian framework the parameters α and β are random variables. These are assigned prior distributions that represent our beliefs about them before considering any empirical data (usually this is also true for σ^2 as this is rarely known

in practice). When the data are complete, and the prior distributions are non-informative relative to the data likelihood, the posterior means of these two parameters are asymptotically normal and coincide, in large samples, with the ordinary least squares estimates obtained in a frequentist setting (also equivalent to their maximum likelihood estimates).

When there is missingness in Y_2 the posterior distributions of α and β can be used to obtain predictions of the missing Y_2 values. Under a MAR mechanism this can be done for subject k, with observed Y_1 and missing Y_2, by obtaining the predictive distribution for the missing value $\widetilde{Y}_{k2}$

$$p(\widetilde{Y}_{k2} \mid Y^O) = \int p(\widetilde{Y}_{k2} \mid \theta, Y_{k1}) p(\theta \mid Y^O) d\theta$$

Here $\theta = (\alpha, \beta)$' are the parameters, Y^O are the observed parts of Y_1 and Y_2, and $\widetilde{Y}_{k2}$ is the missing value to be predicted. $p(\theta \mid Y^O)$ is the posterior distribution for the parameters and $p(\widetilde{Y}_{k2} \mid \theta, Y_{k1})$ is the predictive distribution of the missing Y_2 value given a particular set of parameters θ and the baseline Y_1 value.

6.1.2 Links to Multiple Imputation and classical predictions

The posterior means of the predictive distributions for the missing values coincide with the predicted values $\hat{Y}_{k2} = \hat{\alpha} + \hat{\beta} Y_{k1}$ that can be obtained after fitting a linear regression in the classical setting. Since multiple imputation is essentially a Bayesian procedure for predicting values through the predictive distribution given above, the posterior means also coincide asymptotically with the means of multiply imputed values as the number of imputed datasets, *m,* increases.

6.2 Joint modelling assuming a multivariate distribution for the outcomes

Instead of creating a dependency through the linear predictor (essentially making one the *de*pendent and the other the *in*dependent variable) the outcomes Y_1 and Y_2 can be modelled jointly according to a multivariate distribution.

$$Y_{ij} \sim MVN(\mu, \Sigma) \qquad j=1,2;\ i=1,\ldots,n.$$

Here *MVN* is the multivariate normal density, $\mu = (\mu_1, \mu_2)'$ are the marginal means of Y_1 and Y_2 and Σ is the covariance matrix for Y.

In a Bayesian analysis we need to specify prior distributions for μ and Σ. In WinBUGS this can be done by using a multivariate normal for μ and a Wishart distribution for the precision matrix defined as Σ^{-1}. The Wishart distribution can be treated as a multivariate extension of the chi-squared distribution and is often specified as the prior distribution for precision parameter matrices in WinBUGS (Spiegelhalter et al., 2003).

Having specified appropriate priors for the parameters we may obtain predictive distributions for Y_2 conditional on the data as before. The posterior means of these predictive distributions can be used to estimate $\mu_2=E[Y_2]$ under MAR. Similarly, the posterior means for $\mu = (\mu_1, \mu_2)'$ can be used to obtain valid estimates of $E[Y_1]$ and $E[Y_2]$ under MAR, and will coincide with the estimates derived through the regression formulation above.

Under the regression formulation Y_2 was dependent on Y_1 and could therefore be predicted using Y_1, but there was no feedback mechanism enabling prediction of Y_1 from observed values of Y_2. With the joint model formulation we can model missingness in both Y_1 and Y_2 simultaneously. That is, we are no longer restricted to the situation where only Y_2 is subject to missingness. Under the joint model formulation it is possible to generate predictions of non-monotone missing values. The incomplete outcomes are modelled jointly and explanatory variables can enter the imputation model through the linear predictor.

6.3 Inducing dependency through hierarchical models with exchangeable parameters

Another similarly flexible way of modelling repeated outcomes is through a model of the form

$$Y_{ij} \sim N(s_i, \sigma_e^2) \qquad j=1,2;\ i=1,\ldots,n.$$

where s_i is a subject-specific parameter akin to a random intercept. σ_e^2 is the variance of Y (assuming equal variances at t_1 and t_2) and, conditional on s_i, Y_1 and Y_2 are assumed independent, i.e. $Cov(Y_{i1}, Y_{i2} \mid s_i) = 0$.

Prior distributions are required for the s_i parameters (as well as σ_e^2). We could assume that the s_i are independently distributed by assigning a separate, independent prior to each s_i.

$s_i \sim N(\mu, \sigma_s^2)$, for $i=1,\ldots,n$,

and choosing values for μ and σ_s^2 to ensure flat priors relative to the data likelihood. Alternatively, instead of assuming independent vague priors for the s_i we may let μ and σ_s^2 carry information by modelling these as parameters (known as hyper-parameters), and specifying prior distributions for these also.

This latter alternative recognises that data collected on other subjects in the sample are useful in predicting missing observations from incomplete cases. The individual s_i are estimated using data from the whole sample, not just those belonging to subject i. This is equivalent to the way that a random effects model combines information from all subjects with data from the individual subject to predict subject-specific random effects (known as shrinkage; e.g. Fitzmaurice, Laird & Ware, 2004; Chapter 8). Indeed with this model we obtain results that are exactly equal to those obtained using a random intercepts model in a classical setting. When using vague priors the two approaches are equivalent and yield identical estimates, valid under MAR, of $\mu_2 = E[Y_2]$, σ_e^2 and σ_s^2.

6.4 Bayesian analysis of the POD Study data under a MAR assumption using WinBUGS

Using a hierarchical model (exchangeable parameters)

The following model was fitted to the PODS data.

$$E[Y_{ij}] = \mu_{ij} = b_i + Time_j$$

Here Y_{ij} is the jth response from the ith patient, b_i are the subject-specific means (imposing random intercepts), and $Time_j$ are the freely modelled additional effects of time at month 4, 8, 16 and 28 (i.e. $Time_1=0$).

Using WinBUGS, the data likelihood was then modelled according to $Y_{ij} \sim N(\mu_{ij}, \sigma_e^2)$. Vague normal prior distributions were specified for the parameters $Time_2,\ldots, Time_5$. A hierarchical prior was used for the subject-specific

$$b_i \sim N(\mu_b, \sigma_b^2)$$

with vague priors for the hyper-parameters μ_b and σ_b^2. WinBUGS requires a different parameterisation of the normal distribution using the precision $\tau = 1/\sigma^2$ in place of the variance parameter σ^2. Vague gamma prior distributions were specified for all precision parameters where applicable. The means and standard deviations of the posterior distributions for the parameters are shown in Table 6.1.

Table 6.1. Bayesian analysis of the mean profiles using a hierarchical model

Parameter	Posterior mean	Posterior standard deviation
μ_b	18.66	0.33
$Time_2$	-5.25	0.31
$Time_3$	-4.93	0.32
$Time_4$	-5.16	0.32
$Time_5$	-6.10	0.33
σ_e^2	15.95	-
σ_b^2	18.25	-

These results were compared with the results of an equivalent random intercept model fitted in SAS (Table 6.2).

Table 6.2. Analysis of the mean profiles in SAS using a random effects model.

Parameter	Estimate	Standard error
μ_b	18.66	0.32
$Time_2$	-5.25	0.31
$Time_3$	-4.93	0.32
$Time_4$	-5.16	0.32
$Time_5$	-6.10	0.32
σ_e^2	15.95	-
σ_b^2	18.31	-

As would be expected the results are virtually identical. Using the above parameter estimates, the estimated mean scores under a MAR assumption were 13.73, 13.50, 12.57 at 8, 16 and 28 weeks. At 0 and 4 weeks the estimated means coincided with the observed data means since there were no missing data on these two occasions. These estimates are also in good agreement with the results obtained in Chapter 5 using multiple imputation.

Explicit multivariate normal likelihood

Next as an alternative approach to the analysis of the mean responses in WinBUGS, the outcomes at 4, 8, 16 and 28 weeks were modelled jointly by specifying a multivariate normal distribution.

$Y_{ij} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Here $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ are the marginal means at the four follow-up time points and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33}^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44}^2 \end{bmatrix}$$

is the associated covariance matrix. Vague prior distributions were specified for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma^{-1}}$. The initial scores at 0 weeks were not included in this analysis since these

outcomes were from a truncated distribution and had the potential to affect the multivariate modelling procedure in an untoward manner.

As described in detail in Chapter 5 there were missing data at 8, 16 and 28 weeks. Having fitted the above model to the relevant outcomes, posterior mean estimates were obtained for all of the above parameters (Table 6.3).

Table 6.3. Bayesian analysis of the mean profiles using a multivariate normal data likelihood.

Parameter	Posterior mean	Posterior standard deviation	Parameter	Posterior mean	Posterior standard deviation
μ_1	13.37	0.28	σ^2_{44}	45	3.8
μ_2	13.66	0.31	σ_{12}	25	2.5
μ_3	13.52	0.33	σ_{13}	23	2.5
μ_4	12.52	0.34	σ_{14}	22	2.5
σ^2_{11}	32	2.5	σ_{23}	29	2.9
σ^2_{22}	40	3.2	σ_{24}	28	2.9
σ^2_{33}	43	3.5	σ_{34}	29	3.1

Finally, when pooling the means of the predictive distributions for the missing data with the means of the observed outcomes at 8, 16 and 28 weeks, the following estimates for the average scores were obtained: 13.73 at 8 weeks, 13.56 at 16 weeks and 12.61 at 28 weeks. These results were confirmed in a multiple imputation analysis in SAS including in the imputation model all outcomes from 4, 8, 16 and 28 weeks and using $m=100$ imputed datasets. The means of the imputed data at the three time points subject to missingness were 13.73, 13.55 and 12.61.

6.5 Modelling the non-response mechanism: a simulation exercise

The previous sections focussed on the response processes from incomplete data but ignored the missingness process. In this section we will consider in detail some alternative missingness processes and the practical implications arising when attempting to model these. We will do this through analysis of simulated incomplete data generated from known response and non-response mechanisms in SAS.

6.5.1 Missingness At Random (MAR)

Consider again the scenario from section 6.1: Variable Y is measured on two occasions, t_1 and t_2, resulting in two potentially observed responses, Y_{i1} and Y_{i2}, from the ith subject. In simulation work we let Y have a multivariate normal distribution with mean parameters $\boldsymbol{\mu} = (5\ 6)$' and covariance matrix Σ of dimensions (2×2) with the two diagonal elements equal to 1 and off-diagonal elements equal to 0.5. That is, the marginal means at t_1 and t_2 are 5 and 6 respectively, $Var(Y_{ij}) = 1$, and $Cov(Y_{i1}, Y_{i2}) = 0.5$.

MAR under monotone missingness patterns

In the first instance we let Y_2 be missing in some cases while Y_1 was fully observed. Letting $M_{ij}=1$ when Y_{ij} is missing (and $M_{ij}=0$ otherwise) we denote $P(M_{i2}=1) = \pi_{i2}$, and define the dependency on Y_1 through the following mechanism

$$\log\left(\frac{\pi_2}{1-\pi_2}\right) = \alpha + \beta(Y_1 - 5)$$

Because the covariate is mean centred α has the straightforward interpretation of the overall log odds of missingness in Y_2. β is the increase in the log odds of missingness for each unit increase in Y_1.

We simulated data from 3500 subjects according to the above mechanism with $\alpha = 0$ and $\beta = 0.7$. These translate into modelled probabilities of missingness as shown in Table 6.4. The *complete* (observed and unobserved) data means were $\bar{Y}_1 = 5.03$ and $\bar{Y}_2 = 6.04$. The *observed* data mean at t_2 was $\bar{Y}_2^O = 5.85$; there was 51% missing data on Y_2.

Table 6.4. Probability of missingness in Y_2 as a function of Y_1.

Y_1	1	2	3	4	5	6	7	8	9
$P(M_2 \mid Y_1)$	0.06	0.11	0.20	0.33	0.50	0.67	0.80	0.89	0.94

The data were analysed in WinBUGS using the following random intercepts model for the responses:

$$Y_{ij} \sim N(\mu_{ij},\ \sigma_e^2)$$

$$\mu_{ij} = b_i + Time_j$$

b_i are the subject-specific intercepts and $Time_j$ is the additional effect of the measurement occasion with $Time_1=0$. Vague prior distributions were specified for $Time_2$ and σ_e^{-2}. We specified a hierarchical prior for b_i:

$$b_i \sim N(\mu_b,\ \sigma_b^2)$$

with vague prior distributions on the hyperparameters μ_b and σ_b^{-2}. A separate model was specified for the missing data indicators:

$$M_{i2} \sim Bernoulli(\pi_{i2})$$

$$logit(\pi_{i2}) = \alpha + \beta(Y_{i1} - \bar{Y}_1)$$

Here $\bar{Y}_1$ is the sample mean at t_1. Vague normal priors were specified for α and β. The means and standard deviations of the posterior distributions for the parameters are shown in Table 6.5 (a) ignoring the missing data mechanism and (b) when including the missing data mechanisms in the model.

Table 6.5. Results from analysis with MAR data with a monotone missingness pattern.

(a) Ignoring the non-response mechanism			(b) Joint modelling of the response and non-response mechanisms		
Parameter	Posterior mean	Posterior standard deviation	Parameter	Posterior mean	Posterior standard deviation
μ_b	5.03	0.02	μ_b	5.03	0.02
$Time_2$	1.01	0.02	$Time_2$	1.01	0.02
σ_e^2	0.51	0.02	σ_e^2	0.51	0.02
σ_b^2	0.53	0.02	σ_b^2	0.53	0.02
			α	0.03	0.04
			β	0.77	0.04

Under a MAR mechanism the missingness function is ignorable; the two modelling approaches are equivalent and agree with the complete data means. The posterior

means for the parameters governing the non-response mechanism were close to the true parameter values behind the simulated data of $\alpha = 0$ and $\beta = 0.7$.

Multiple incomplete variables with monotone missingness patterns

The scenario above can be extended to include monotone missingness in two or more variables. Much of the literature on missing data is concerned with modelling drop-out mechanisms rather than intermittent missing data of a non-monotone nature (wave non-response). There is often a preference to avoid modelling non-monotone missingness patterns under a MAR assumption. One way to avoid this is to assume a MCAR mechanism for wave non-response, but to treat drop-out under a MAR assumption (e.g. Carrigan et al., 2007). At first this may seem an unrealistic simplification, however it is often difficult to think of a plausible missing data process that causes a non-monotone missingness pattern under a MAR mechanism.

MAR under a non-monotone mechanism

Recall that under a MAR mechanism the probability distribution for the missingness indicators may depend on the observed elements of the response vector, but cannot depend on the unobserved elements.

With monotone missing data as caused by drop-out this leads to a relatively straightforward process whereby the probability of drop-out is determined by (all or some of) the previously observed responses. We can imagine a simple process where the probability of drop-out at time j is random conditional on the response at time $j-1$.

With non-monotone data it is not as straightforward. To see why, consider a situation where the probability of missingness at time j is determined by all or some of the observed elements of Y. Since with non-monotone data none of the time points are uniformly more observed than others, and because MAR missingness cannot depend on unobserved values, the missingness process at any particular time point is necessarily dependent on which elements of Y are observed. In other words, the probability of missingness in Y_2 can depend on Y_1 only if Y_1 is observed, but will have to depend on something else if Y_1 is not observed.

When forced to think about the functional forms of the missing data process under a MAR mechanism in non-monotone data it is often apparent that such processes are possible only in rather contrived constructions. Conveniently the non-response process is often not of direct interest, and one can proceed to estimate the response model under a MAR assumption without having to consider the form of the missing data process. However some authors have pointed out that statistical convenience alone does not justify the reliance on implausible assumptions and that more plausible processes under a MNAR mechanism should be considered instead (e.g. Robins & Gill, 1997).

Simulation exercise

Using the same bivariate normal data example from before we simulated non-monotone missing data that were MAR. We generated missing data under a process where

$$P(M_1=1) = \pi_1 = 0.3$$

and depending on M_1

$$\log\left(\frac{\pi_2}{1-\pi_2}\right) = \alpha + \beta(Y_1 - 5) \quad , \qquad \text{for } M_1=0$$

or

$$\log\left(\frac{\pi_2}{1-\pi_2}\right) = \alpha \quad , \qquad \text{for } M_1=1$$

Again we simulated data from 3,500 subjects. The complete data means were: $\bar{Y}_1 = 5.00$; $\bar{Y}_2 = 6.02$. The observed data means were: $\bar{Y}_1^O = 4.99$; $\bar{Y}_2^O = 5.94$. There was 51% missingness on Y_2 and 32% on Y_1. The results from modelling these incomplete data in WinBUGS using the same model specification as before are shown in Table 6.6.

Table 6.6. Results from analysis with non-monotone missing data from a MAR mechanism (n=3,500).

Ignoring the non-response mechanism		
Parameter	Posterior mean	Posterior standard deviation
μ_b	5.00	0.02
$Time_2$	1.03	0.03
σ_e^2	0.53	0.02
σ_b^2	0.46	0.03

6.6 Missingness Not At Random (MNAR)

Again using the bivariate normal data example from above we simulated 3500 sets of data. Missingness was then induced in Y_1 and Y_2 with probabilities π_1 and π_2 governed by the following mechanism

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_1$$

$$\log\left(\frac{\pi_2}{1-\pi_2}\right) = \alpha_2 + \beta(Y_1 - 5)$$

with $\alpha_1 = -0.85$ (i.e. a constant missingness probability of $\pi_1 = 0.3$), $\alpha_2 = 0$ and $\beta = 0.7$. (Notice how missingness in Y_2 now depends on the potentially unobserved value of Y_1.) The *complete* (observed and unobserved) data means were $\bar{Y}_1 = 5.03$ and $\bar{Y}_2 = 6.04$. The *observed* data means were $\bar{Y}_1^O = 5.03$ and $\bar{Y}_2^O = 5.86$. There was 30% missingness on Y_1 and 49% on Y_2.

These data were again analysed in WinBUGS using the same hierarchical model from before for the response mechanism. The results from fitting this model ignoring the non-response mechanism are shown in Table 6.7.

Table 6.7. Results from analysis with MNAR data, ignoring the non-response mechanism (N=3,500).

(a) Random effects model

Parameter	Posterior mean	Posterior standard deviation
μ_b	5.02	0.02
$Time_2$	0.95	0.03
σ_e^2	0.51	-
σ_b^2	0.53	-

(b) specifying a multivariate normal likelihood

Parameter	Posterior mean	Posterior standard deviation
μ_1	5.02	0.02
μ_2	5.97	0.02
σ_{11}^2	1.05	0.03
σ_{22}^2	1.03	0.04
σ_{12}^2	0.51	0.03

The results obtained from fitting this model did not agree entirely with the complete data means. In particular it seems that $Time_2$, or equivalently $E[Y_2]$ was underestimated, albeit to a very small degree. Since we have previously encountered disagreements between the multilevel modelling approach and the model specifying an explicit multivariate likelihood, it was important to rule out the modelling approach as a reason for the discrepancy. For this reason the data were modelled a second time specifying a bivariate normal distribution for the Y values with mean vector $\mu=(\mu_1, \mu_2)$' and $Cov(Y_j,Y_k)=\sigma_{jk}^2$ (Table 6.7(b)). The two approaches agree as we would expect with normally distributed data, but do not agree with the complete data means. We therefore repeated the analysis but on a larger dataset using 10,500 sets of observations (Table 6.8). The *complete* (observed and unobserved) data means were $\bar{Y}_1 = 5.00$ and $\bar{Y}_2 = 6.01$. The *observed* data means were $\bar{Y}_1^O = 5.00$ and $\bar{Y}_2^O = 5.86$. As before there was 30% missingness on Y_1 and 49% on Y_2.

Table 6.8. Results from analysis with MNAR data ignoring the non-response mechanism (N=10,500).

Parameter	Posterior mean	Posterior standard deviation
μ_1	4.99	0.01
μ_2	5.97	0.01
σ_{11}^2	1.02	0.02
σ_{22}^2	0.98	0.02
σ_{12}^2	0.48	0.02

Again $E[Y_2]$ was underestimated compared with the complete data mean. In conclusion, although missingness in Y_2 did not depend on the value of the unobserved Y_2 value itself, modelling the response mechanism alone is not sufficient for unbiased estimation of $E[Y_2]$ even under this type of MNAR mechanism.

6.7 Summary

This chapter provided an account of early exploratory analysis with simple practical examples conducted by the author to gain experience necessary for subsequent analyses presented in the thesis. The chapter was introduced with an exploration into some alternative ways in which correlated data might be modelled using WinBUGS. Some well-known ideas were illustrated using practical examples. We then applied a few simple analyses with the incomplete PODS data comparing results from analyses with SAS and WinBUGS to ensure it was possible to obtain results as expected, and to gain familiarity with Bayesian analysis with incomplete data in WinBUGS. We then simulated bivariate data in SAS and induced missingness in the data through simple non-response mechanisms. The incomplete data were simulated under monotone MAR, non-monotone MAR and MNAR mechanisms. We experimented with alternative model fits to the simulated data in WinBUGS (including joint model specifications for the data and missingness mechanisms) and obtained results that were consistent with theory. These reassuring results illustrated that WinBUGS is a powerful, practicable tool that can be used for fitting models to the incomplete data that are the subject of subsequent chapters.

7 ANALYSIS WITH THE SCREENING DATA USING SAS AND WINBUGS

The previous chapter was concerned with the practicalities of modelling incomplete longitudinal data in WinBUGS with the aim of building up a practical understanding of ways to approach the modelling exercise. In this chapter we shall return to the core problem of how routinely collected repeated patient outcomes might be used to address research questions.

The routine collection of patient outcomes may be of direct benefit to the patient in terms of the care they receive by facilitating communication and enabling clinicians to react to significant worsening of symptoms etc. On the other hand it is less clear how, at the aggregate level, such data may be used for research. Often the amount of routinely collected data exceeds by far what researchers can hope to collect in a purpose-designed clinical study since recruitment of patients into research projects is notoriously difficult and expensive. An interesting question is therefore whether small, but purpose-designed studies are better equipped to address research aims than large-scale routinely collected data. The POD Study, which was the subject of Chapter 5, was a purpose-designed, clinical study executed to address a set of pre-defined research questions. In this chapter we ask if the aims of the POD Study could have been addressed using data that were routinely collected by the Symptom Monitoring Service. Specifically, we aim to replicate the POD Study analysis of mean distress levels and prevalence of distress after one, two, four and seven months among patients identified with distress (HADS score≥15) at an initial (qualifying) clinic appointment, and secondly, in the same cohort, the associations with persistent distress at seven months.

7.1 The screening service

The Symptom Monitoring Service which was described in detail in section 4.1 routinely administered a questionnaire containing the HADS to patients when they attended clinics. Patients with frequent clinic appointments were consequently observed more often than patients with infrequent appointments.

The service also recorded a reason when patients with appointments failed to complete the screening: some patients declined the offer of screening, others were unable to complete screening due to very poor health or communication difficulties. Sometimes patients were simply missed by the screening service, usually because they were taken straight to their consultation before being screened. The service also did not ask patients to repeat the screening if their appointments were less than four weeks apart. The limitations to the data were therefore considerable, partly because of missing data from patients who attended the clinics, and partly because of the opportunistic data collection method whereby patients were sampled only when they attended medical appointments.

7.2 Derivation of the analysis sample

During its three year period of operation the depression screening service logged over 100,000 appointments from more than 30,000 patients (Figure 7.1). Of these 24,260 patients completed the screening questionnaire at least once and gave permission for their questionnaire data and data about their cancer to be used for research in anonymised form. It was important that the sample selected for the current analysis corresponded to the POD Study sample as this would allow for a direct comparison. To be eligible for inclusion in the analysis set patients therefore had to:

(1) have scored 15 or more on the HADS when screened during a clinic appointment
(2) not meet criteria for major depression in the subsequent telephone interview
(3) have a prognosis of at least 12 months as estimated by their cancer specialist
(4) meet the first three criteria before October 2010 to allow for a reasonable period for follow-up data to accumulate.

Patients who met the above eligibility criteria on more than one occasion were included in the analysis on the basis of the first occasion the criteria were satisfied. (We shall refer to this as the baseline visit.)

We also obtained NHS Caldicott approval and approval from the NHS Scotland Privacy Advisory Committee to link patients' questionnaire data with data about their cancer held centrally by the Scottish Cancer Registry. The Registry is managed by NHS Scotland Information Services Division (ISD) who linked our data to the oncology records and prepared a merged database in anonymised form. We identified 2,180 cases that satisfied the eligibility criteria. Of these ISD had linked oncology records to 2,149 (99%) cases that were subsequently included in the analysis.

7.3 Characteristics of the analysis sample

The basic demographic and clinical characteristics of the analysis sample are presented in Table 7.1. The mean age was 62.3 years (SD: 11.9 years). The majority of patients included were female (81%), had breast (47%), gynaecological (18%) or bowel cancer (17%), and were treated with curative intent (81%). The median time since diagnosis was 1.4 years (IQR: 0.4 to 3.9 years). Most patients scored 15 or just over on their baseline HADS.

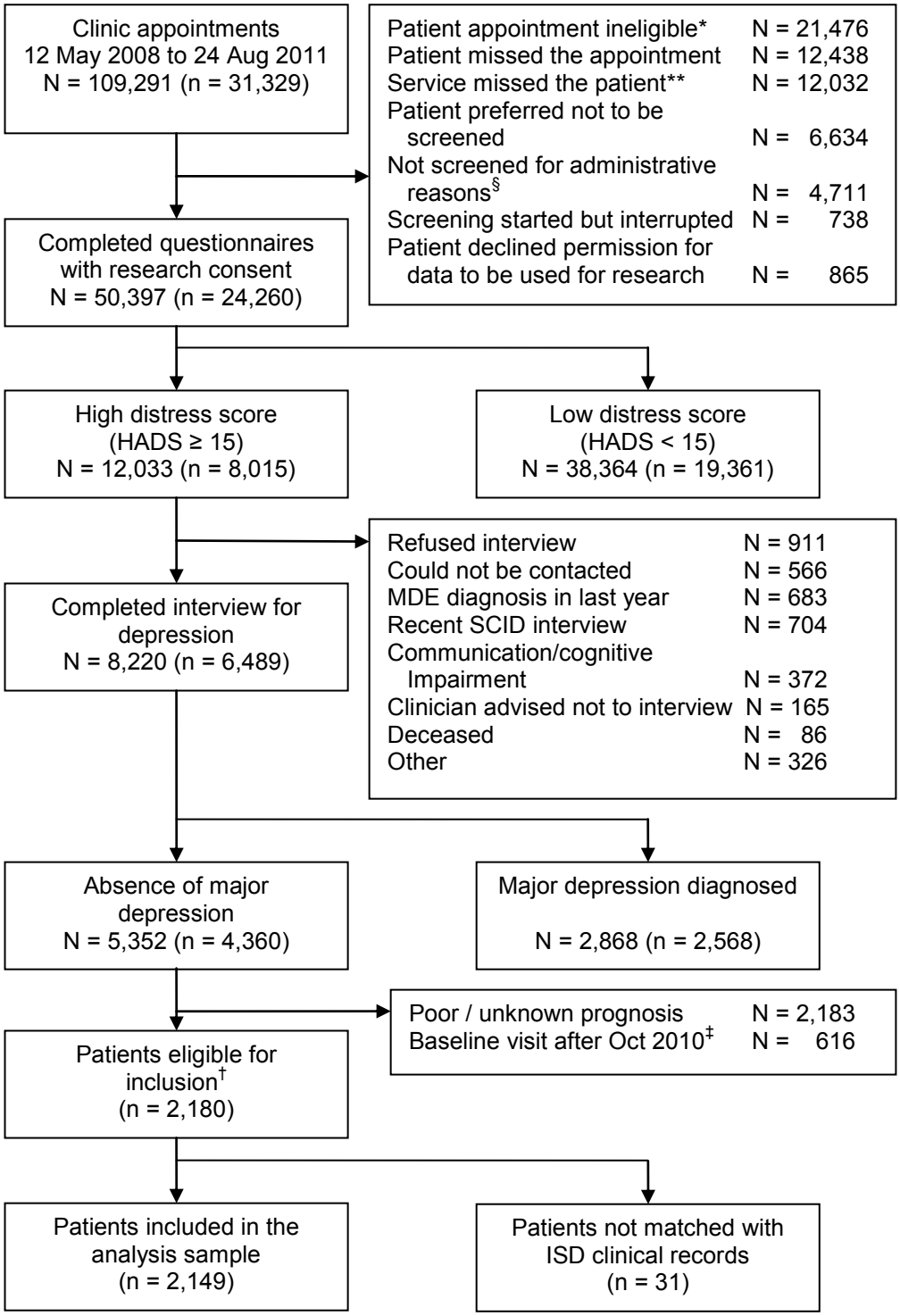


Figure 7.1. Derivation of the analysis sample. *N* indicates number of appointments. *n* in parentheses indicates number of patients. † When patients met eligibility criteria more than once we used the first occasion as baseline in the analysis. * Patient screened very recently, under 18 years of age, too unwell or distressed, had cognitive or visual impairment, hearing or language problems, clinician advised against screening, patient's diagnosis unclear, including other reasons. ** Typically patients were missed in the waiting area because they were called in for their appointment before the screening service could approach them. § Reasons include cancellations of scheduled clinics, patient appointment changed to a different clinic. ‡ To ensure a reasonable follow-up period we excluded patients whose baseline appointment occurred after 1st October 2010.

Table 7.1. Demographic and clinical characteristics of the analysis sample at baseline.

Total	2149 (100%)
Gender	
Female	1745 (81%)
Male	404 (19%)
Age (years)	
mean (SD) min – max	62.3 (11.9) 20 – 92
median (IQR)	63 (54 – 71)
Age group	
< 50 years	328 (15%)
50 to 64 years	858 (40%)
≥ 65 years	963 (45%)
Health board*	
Lothian	627 (29%)
Glasgow	506 (24%)
Argyll & Clyde	314 (15%)
Lanarkshire	247 (11%)
Forth Valley	147 (7%)
Fife	106 (5%)
Tayside	97 (5%)
Other†	105 (5%)
Cancer type	
Bowel	370 (17%)
Breast	1003 (47%)
Genitourinary	185 (9%)
Gynaecological	386 (18%)
Other	205 (10%)
Time since diagnosis (years)	
mean (SD) min; max	2.9 (3.8) 0.0 to 30.1
median (IQR)	1.4 (0.4 – 3.9)
Treatments started in the past 2 months ‡	
Surgery	251 (12%)
Radiotherapy	34 (2%)
Chemotherapy	100 (5%)
Care objective §	
Curative	1473 (81%)
Palliative	351 (19%)
Deprivation SIMD quintile score**	
1	480 (22%)
2	450 (21%)
3	370 (17%)
4	373 (17%)
5	476 (22%)
HADS score at baseline	
mean (SD) min; max	18.8 (3.7) 15 – 38
median (IQR)	18 (16 – 21)

Data are number (%) unless otherwise indicated. * Health board where patient was resident when cancer was registered. ** Scottish Index of Multiple Deprivation quintile score: 1=most deprived, 5=least deprived. † Ayrshire and Arran, Borders, Dumfries & Galloway, Grampian and Western Isles. ‡ Variables incompletely observed: surgery N=2056, radiotherapy N=2000, chemotherapy N=2048. § Care objective incompletely observed: N=1824.

7.4 Four time points

The sample of 2,149 patients had more than 8,000 scheduled appointments logged in the screening database following the initial (baseline) clinic visit. Most patients had just a few appointments while a small number of patients had numerous.

In the POD Study, participants were followed up for a period of seven months after scoring high on the HADS at the baseline clinic visit; HADS measurements were obtained again at 1, 2, 4 and 7 months. The clinic appointments in the screening data did not naturally follow this temporal structure of the PODS data. We therefore divided the screening data into four bins where observations obtained between Day 1 and Day 52 were classed as Month 1 data, observations between Day 53 and Day 98 were classed as Month 2 data, observations between Day 99 and Day 177 were classed as Month 4 data and observations between Day 178 and Day 277 were classed as Month 7 data.

We required just one observation per patient per time point. Where multiple observations were available within a single time bin, we selected the observation closest in time to the median time point observed in the PODS data for the relevant time point. Thus, for each patient we selected observations nearest in time to 34, 70, 127 and 227 days for each of the nominal times at 1, 2, 4 and 7 months. The boundaries separating each of the four bins at 52, 98 and 177 days were the midpoints between the median time points and were chosen on the basis of providing good separation between the time points in the PODS data (Figure 7.2). Although somewhat arbitrary, the upper limit of 277 days for the Month 7 window ensured a symmetrical window about the median time for that time bin. The derived time windows fitted around the observations included for analysis from the screening data are illustrated in Figure 7.3.

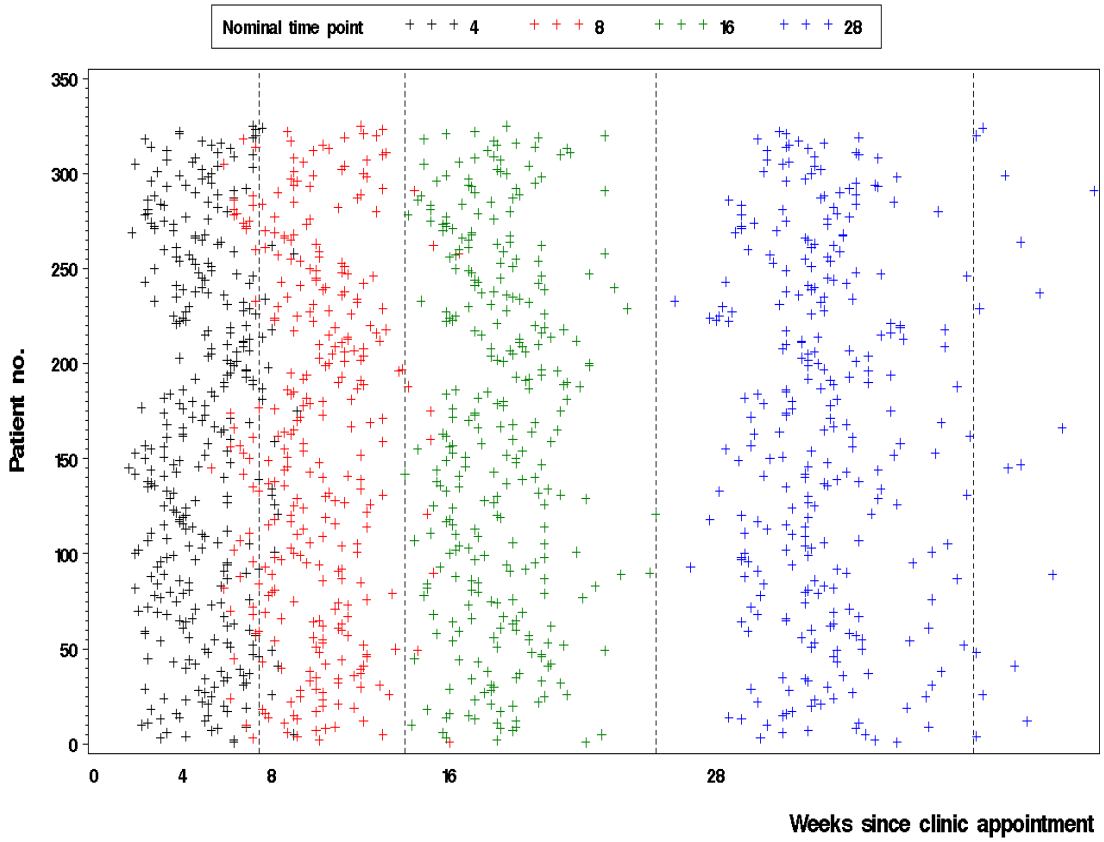


Figure 7.2. Timing of questionnaires collected in PODS.

It is possible that a bias was introduced in prioritising patient appointments where the HADS was completed instead of appointments closest in time to the median time point, whether or not the HADS was observed. However the effect arising from such a bias is likely to have also been present in the POD Study where patients had the opportunity to complete questionnaires later if they felt unwell when first approached.

7.5 Exploratory analysis

All 2,149 patients necessarily provided data at the baseline visit. Table 7.2 lists the number of patients with appointments during follow-up and the HADS measurements obtained at those appointments. (The lower proportion of patients with observed data at 1 month relative to the number of patients with appointments was due to a change in procedures whereby patients screened after October 2009 were not asked to complete a screening questionnaire in the clinic if they had been screened once already in the past four weeks.)

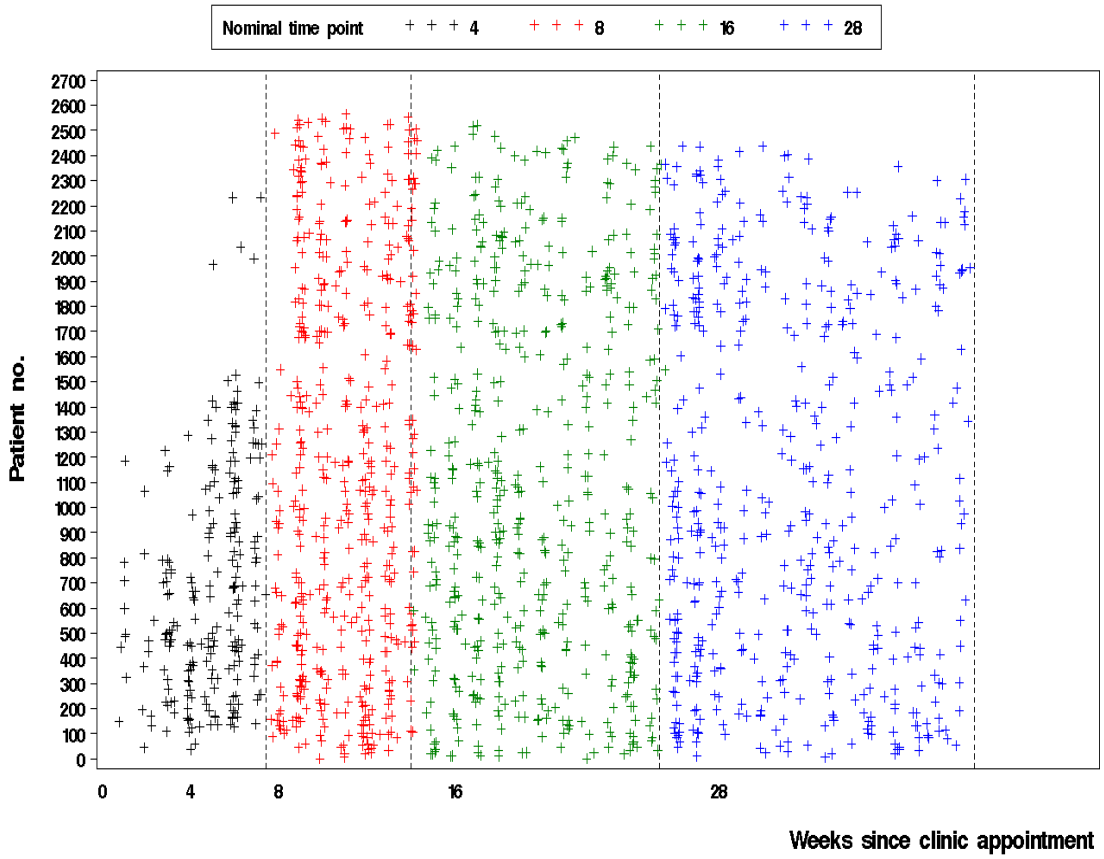


Figure 7.3. Timing of observations included for analysis from the screening data.

Table 7.2. Appointments and data completeness.

	Patients with appointments	Patients with observed HADS
Baseline	2149 (100%)	2149 (100%)
1 month	632 (29%)	226 (11%)
2 months	738 (34%)	437 (20%)
4 months	814 (38%)	499 (23%)
7 months	825 (38%)	490 (23%)

Data are number (%).

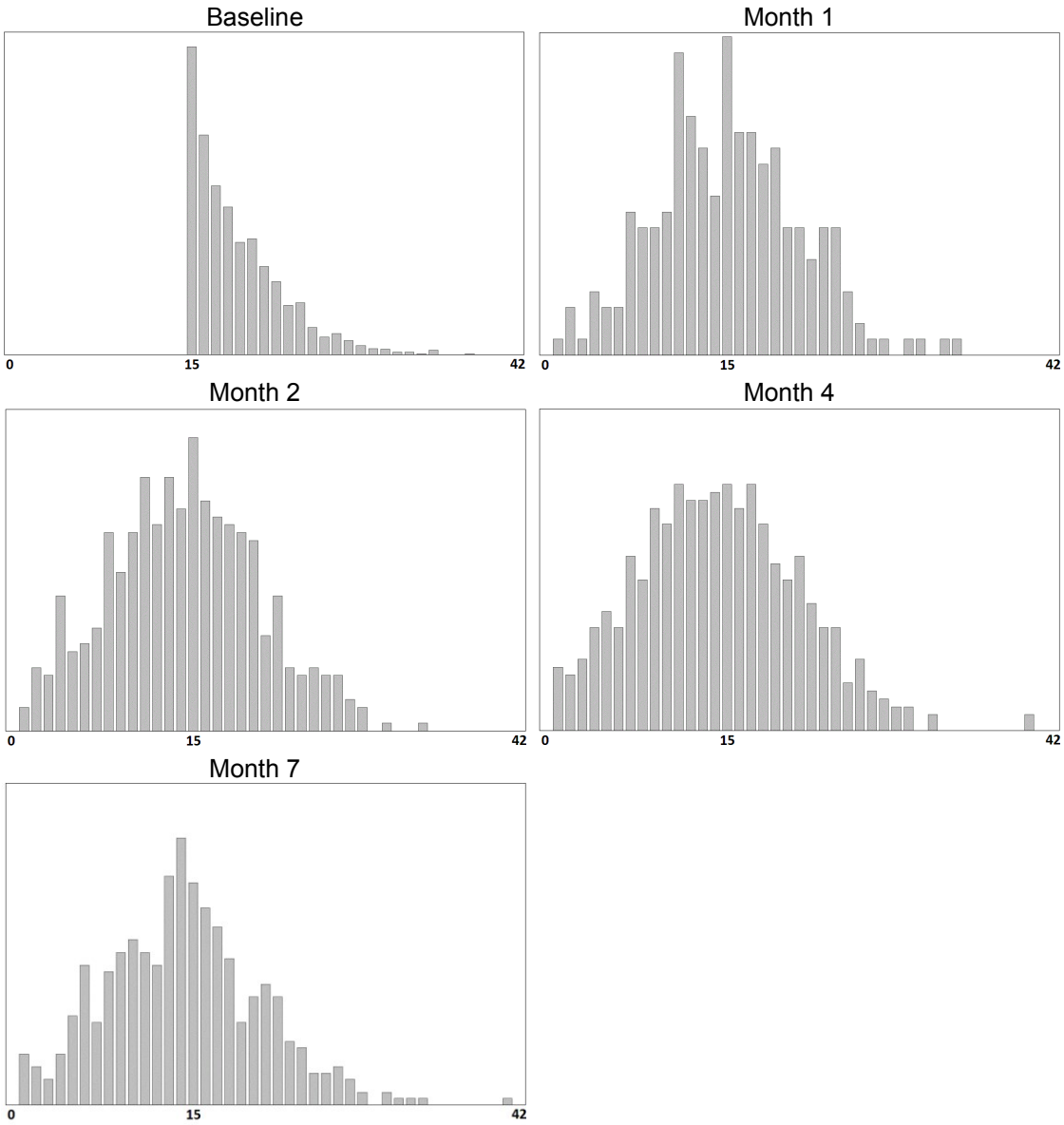


Figure 7.4. Marginal distributions of the observed HADS scores obtained at baseline and during follow-up

The marginal distributions of the observed HADS scores are presented in Figure 7.4. The distribution was truncated at baseline, but quickly became more symmetrical at subsequent time points. The variance appears to remain fairly constant during follow-up.

7.5.1 Maximum likelihood estimation of the covariance matrix

We fitted a multivariate normal linear model to the HADS scores allowing for separate intercepts at 1, 2, 4 and 7 months, and estimated the unstructured covariance

matrix using maximum likelihood estimation (Table 7.3). The associated correlation matrix is shown in table 7.4.

Table 7.3. Maximum likelihood estimates of the variance-covariance matrix.

	Month 1	Month 2	Month 4	Month 7
Month 1	35.9			
Month 2	22.0	38.4		
Month 4	23.9	27.4	45.7	
Month 7	20.9	25.1	31.4	42.3

Table 7.4. ML estimates of the correlation matrix.

	Month 1	Month 2	Month 4	Month 7
Month 1	1			
Month 2	0.59	1		
Month 4	0.59	0.65	1	
Month 7	0.54	0.62	0.71	1

We also fitted a reduced model with a compound symmetry covariance instead of the unstructured pattern thereby reducing the number of covariance parameters from 10 to just 2. The resultant (constant) variance estimate was 41.4, the covariance was 26.4, and the correlation was therefore 0.64. There was no statistically significant evidence that the full model provided a better fit (likelihood ratio test statistic = 14.6, d.f. = 8, p=0.067).

7.5.2 Why the missing data?

There was a considerable amount of missing data, although this was mostly due to lack of patient appointments within the four time windows. The main reasons for missing data from patients with clinic appointments are listed in Table 7.5.

The majority (51%) of the sample did not provide data at any of the follow-up time points. Twenty-eight percent provided data at one time point only, 13% were observed on only two occasions, 6% on three occasions only, and just 1% were observed on all four follow-up occasions (Table 7.6).

Table 7.5. Reasons for missing data in patients with clinic appointments.

	Week 4	Week 8	Week 16	Week 28
Total missing	406 (100%)	301 (100%)	315 (100%)	335 (100%)
Recently screened	271 (67%)	58 (14%)	56 (14%)	24 (6%)
Patient cancelled/did not attend appointment	37 (9%)	44 (11%)	49 (12%)	67 (17%)
Patient was missed by the screening service	37 (9%)	92 (23%)	86 (21%)	84 (21%)
Patient refused screening	34 (8%)	58 (14%)	65 (16%)	86 (21%)
Other reasons	27 (7%)	49 (12%)	59 (15%)	74 (18%)

Other reasons included patient visually impaired, too ill or too upset, patient's clinician advised against screening and administrative problems.

Table 7.6. Patterns of missingness over the four follow-up time points.

Pattern no.	Week 4	Week 8	Week 16	Week 28	Number (%) of patients with the indicated pattern
1	M	M	M	M	1106 (51%)
2	M	M	O	M	197 (9%)
3	M	M	M	O	195 (9%)
4	M	O	M	M	148 (7%)
5	M	M	O	O	82 (4%)
6	M	O	M	O	72 (3%)
7	O	M	M	M	67 (3%)
8	M	O	O	O	65 (3%)
9	M	O	O	M	58 (3%)
10	O	O	M	M	34 (2%)
11	O	M	O	M	28 (1%)
12	O	O	O	O	24 (1%)
13	O	M	O	O	24 (1%)
14	O	O	O	M	23 (1%)
15	O	O	M	O	13 (1%)
16	O	M	M	O	13 (1%)
Total					2149 (100%)

Note: O (M) indicates that the data point was observed (missed). E.g. a patient with missingness pattern 2 provided data at each of four weeks, eight weeks and 16 weeks, but not at 28 weeks.

7.5.3 Means by missingness patterns

To determine if an obvious relationship existed between the observed distress scores and patients' missingness patterns we plotted the mean distress scores over time separately for each missingness pattern. There were no obvious indicators that patients with less observed data had scored differently on those occasions when their HADS scores were observed than patients with more observed data. To see if a better discriminator could be found in the appointment patterns, we also plotted the mean

distress scores over time, separately for each appointment pattern in the sample. However there did not seem to be any obvious relationship between the level of distress among the available scores and patients' appointment patterns. It seemed there might be other factors, not captured in patients' appointment and missing data patterns, contributing to variations in patients' distress scores.

7.6 The response model

Our goal was to fit models that could predict the missing observations. Taking a principled approach, we hoped to use bits of observed information about the patients to account for variation in the distress scores thereby making the missing data MAR. In that sense a MAR mechanism is a property, not of the data itself, but of the model used to fit the data.

We collected data on the following two stochastic variables: Y, the vector of HADS responses and R, the vector of indicator variables. R takes the value *1* when the associated Y is observed, and *0* otherwise. The covariate data X are treated as completely observed constants for now. The joint distribution for the data from patient i is then

$$f(Y_i, R_i \mid \theta, \psi, X_i)$$

where θ and ψ are the parameters governing the response and the missingness processes respectively. The joint distribution can be partitioned into

$$f(Y_i \mid \theta, X_i)\, f(R_i \mid Y_i, \psi, X_i)$$

which is the selection model decomposition. We are interested in the complete-data distribution for the responses, $f(Y_i \mid \theta, X_i)$; the missing data process is mostly a nuisance. We assume initially that the missingness was governed by a MAR process and, by definition, that the missingness process was independent of the unobserved scores conditional on covariates X.

$$f(R_i \mid Y_i, \psi, X_i) = f(R_i \mid Y_i^O, Y_i^M, \psi, X_i) = f(R_i \mid Y_i^O, \psi, X_i)$$

(Here Y_i^O and Y_i^M are the observed and unobserved segments of Y_i respectively). With the right choice of X we would therefore be able to estimate θ by modelling only the data likelihood for Y while treating the contribution from the missingness process as a constant.

We fitted the below hierarchical model to the HADS responses at 1, 2, 4 and 7 months which were denoted by $Y_{i1},...,Y_{i4}$.

$$Y_{ij} \sim N(\mu_{ij}, \sigma_e^2)$$

$$\mu_{ij} = \alpha_j + X_i^T\beta + b_i$$

$$b_i \sim N(0, \sigma_s^2)$$

Y_{ij} is the normally distributed HADS score from patient i ($i=1,...,2149$) at time j ($j=1,...,4$) with mean μ_{ij} and variance σ_e^2. The α_j are the four freely varying intercepts over time, X_i is the vector of time-invariant covariates from patient i with β the associated regression coefficients. b_i are the independent, normally distributed subject-specific random intercepts.

This model assumes a multivariate normal distribution for the repeated HADS scores at follow-up. The compound symmetry imposed by the multi-level model restricts the variances to be equal at all four time points, and likewise for the covariances $Cov(Y_{ij}, Y_{ik}) = \rho_0$ for all $j \neq k$, so that the correlation between any pair of scores from the same individual is assumed constant regardless of the separation in time between the measurements. These restrictions are arguably unlikely to hold entirely. We chose to impose these in the name of parsimony, to maintain as simple a model as is justifiable. Our preliminary analysis in section 7.5 suggested that these are not unreasonable assumptions.

We fitted the model using Bayesian methods. This required that we specify prior distributions for the stochastic parameters. We specified a multivariate normal distribution for the vector of regression coefficients with zero mean vector and diagonal precision matrix with all entries equal to 1/1000 (WinBUGS parameterises the normal distribution in terms of the mean and precision which is defined as the reciprocal of the variance). We also specified prior gamma distributions for $1/\sigma_e^2$, and for the hyper-parameter $1/\sigma_s^2$. To ensure that inferences about the parameters were based predominantly upon the observed data, the prior distributions were all parameterised so as to be flat relative to the data likelihood.

7.7 The covariate model

The covariates included in the linear predictor were those of the available variables that were either found in exploratory analyses to be associated with the responses or considered important on clinical grounds. The covariates included were: gender, age group (<50 years; 50-64 years; ≥65 years), years since diagnosis (<1 year; 1-5 years; >5 years), cancer type (breast; genitourinary; gynae; lower gastrointestinal; benign; other), therapeutic objective (curative; palliative), surgery, chemotherapy or radiotherapy started in the two months prior to baseline screening (yes; no), patient's residing health board (one of nine, smaller health boards were pooled together), Scottish Index of Multiple Deprivation (SIMD) quintile score, and whether the patient was still alive one year after the follow-up period. Finally, the baseline HADS score (range 15-42) was included as a continuous covariate. The covariates were mostly observed except for the treatment variables (surgery, 4% missing; radiotherapy, 7% missing; chemotherapy, 5% missing), and therapeutic objective (15% missing).

In section 7.6 we noted that the joint distribution of the data could be decomposed in the following way

$$f(Y_i, R_i \mid \theta, \psi, X_i) = f(Y_i \mid \theta, X_i)\, f(R_i \mid Y_i, \psi, X_i)$$

Under a MAR mechanism the last factor on the right, the missingness process, can be treated as a constant in the likelihood expression and can effectively be ignored (that is unless the missingness process is itself of interest). This leaves us with the task of modelling $f(Y_i \mid \theta, X_i)$, the distribution of the HADS scores conditional on the covariate data and the model parameters.

Because the covariate data also are subject to missingness it is helpful at this point to divide the vector of covariates X into two parts. The first part, W, consists of the completely observed elements of X and are thought of as non-stochastic constants. The second part, Q, consists of the incompletely observed elements of X; the covariates subject to missingness. In contrast to the completely observed covariates the elements of Q will be treated in the analysis as random variables. The joint distribution of the random variables (omitting the missingness indicators) can then be factorised in the following way

$$f(Y_i, Q_i \mid W_i, \theta, \gamma) = f(Y_i \mid Q_i, W_i, \theta)\, f(Q_i \mid W_i', \gamma)$$

where θ are the parameters governing the response process as before, γ are the parameters governing the distribution for Q_i, the incompletely observed covariates, and W_i' is a subset of W_i predictive of the missing Q_i. (The elements of W_i' were: years since diagnosis, cancer type and patient's residing health board). We have assumed that the incomplete covariates are MAR conditional on W_i'.

We specified the distribution for the responses as in section 7.6. Letting Q_{ik} denote the kth covariate in the Q vector from patient i we specify the joint distribution for the four incomplete binary covariates Q_1, Q_2, Q_3 and Q_4 (with data on surgery, chemotherapy, radiotherapy and curative intent respectively) through a series of chained logistic regressions:

$$Q_{ik} \sim Bernoulli(\pi_{ik}) \quad , \qquad k=1,\ldots,4$$

$$logit(\pi_{ik}) = \delta_k + W_i'^{T}\gamma_k + \widetilde{Q}_{ik}^{\,T}\lambda_k$$

Here δ_k are the intercepts and ${W_i}'^{T}\gamma_k$ are the effects of the predictor variables on the dependent variable, the incomplete covariate being modelled. The last term of the linear predictor above consists of

$$\widetilde{Q}_{ik}{}^{T} = \{Q_{i1}, \ldots, Q_{i4}\} \backslash \{Q_{ik}\}$$

with associated regression coefficients

$$\lambda_k = \{\lambda_{k1}, \ldots, \lambda_{k4}\} \backslash \{\lambda_{kk}\}.$$

and amounts to the effects of the three elements of Q not being modelled as the dependent variable. That is $\widetilde{Q}_{ik}{}^{T}\lambda_k = \sum_{(l \neq k)} \lambda_{kl} Q_{il}$.

The four binary incomplete covariates were modelled in a series of four linked logistic regressions, each modelling the conditional probability of the covariate being present conditional on all other variables in the model, including the other three jointly modelled covariates. This somewhat convoluted setup of chained models is used to account for dependency between the elements of Q and requires iterative numerical estimation methods. We were able to use WinBUGS to fit the model for the incomplete covariate data jointly with the response model of section 7.6. The underlying assumption is that the simulated draws from the chained conditional distributions amount to the true joint distribution for the four binary covariates. Although the theoretical justification for this approach is lacking, in practice results with Multiple Imputation using Chained Equations (MICE, van Buuren & Oudshoorn, 2000) have been broadly promising, and the method is increasingly regarded as an acceptable approach for imputing non-monotone missing data from arbitrary distributions (van Buuren, 2007, Lee & Carlin, 2010; White, Royston & Wood, 2011).

7.8 Auxiliary data

Many of the patients included in the analysis had either no or only very few observed HADS scores during the seven months of follow-up. However a considerable proportion of these patients had observed HADS data from screening episodes completed either before the baseline visit or after the follow-up period had ended.

Patients with very little observed data during follow-up contributed for obvious reasons with less data to the analysis than patients with more observed data. If these patients were systematically different from the rest of the sample in some way that affected their (unobserved) HADS scores, the omission of data from these patients would be causing bias.

To increase the amount of information available in the analysis from such patients we extended the response model to include data from an auxiliary time point collected outside the study period. It was expected that the inclusion of these auxiliary scores would improve estimation of b_i, the random subject effects, particularly in patients with no observed response data during follow-up whose predicted scores otherwise depended purely on covariate data and shrinkage effects.

We experimented with the number of auxiliary data points to include from each patient and with the timing of them. The inclusion of an early auxiliary data point obtained before the baseline screening episode was problematic for two reasons: Firstly the effects of the covariates on the distress scores were assumed to be constant over time, but with increasing gaps in the time between the first and the last measurements this assumption became increasingly unlikely (and for some of the covariates, such as treatments started in the two months prior to baseline screening, this was clearly not the case). Secondly the estimated common correlation of repeated scores imposed by the hierarchical model was considerably lower when also including responses from before the baseline screening. Eventually we included one auxiliary time point from each patient where available within the first year of the follow-up period ending. There were 587 patients with auxiliary data (27% of all

patients in the sample), 226 of whom contributed with no data during follow-up. The median time since the baseline visit was 11.9 months (IQR: 10.5 to 13.5 months).

The model specified in section 7.6 was thus extended to accommodate the additional time point simply by letting *j* range from one to five instead of four.

7.9 Conditioning on missingness reason

We learned in section 7.5 that there was no discernible association between patients' average distress scores and patterns of missingness, at least not on those occasions when their scores had been observed. However, further analysis suggested that certain types of missingness were associated with variations in patients' observed distress scores. Patients who had refused the option of screening on one or more occasions (the Refusers) had higher distress scores on average when their scores were observed, compared with those who had never refused screening. On the other hand, patients who were excused from screening because they had already been screened at a clinic appointment in the preceding four weeks (the Excused) scored lower on average on those occasions when their scores were observed.

It is reasonable to assume that the value of the *unobserved* scores had they been observed would likewise have been higher in patients who were Refusers and lower in patients who were Excused. This does not violate the MAR assumption since the missingness process can still be independent of the unobserved responses *conditional* on the observed responses (and covariate data). But what exactly is the nature of the relationship between the missing data indicator and the unobserved score, between being a Refuser, say, and the unobserved HADS score? In its current form the model specified in section 7.6 (and extended in 7.7 and 7.8) imposes a dependency between being a Refuser and the unobserved response only via the correlation between repeated responses (and possibly through associations with covariates). But suppose Refusers have some inherent characteristic that affects their HADS score by an amount, δ, regardless of the score being observed or not.

Observed scores from the Refusers would then naturally be an average of δ higher than scores from patients who never refused screening. But the joint modelling approach would only mediate part of the δ effect onto the unobserved scores unless the correlation between repeated scores is perfect, i.e. $\rho=1$. Therefore, if we believe the hypothesis that patients who occasionally refuse screening share a trait that makes them likely to be more distressed on average than patients who never refuse screening, then we ought to condition the responses on the presence of this trait by including it as a covariate in the model. Doing so will ensure that both observed and unobserved scores from patients who are Refusers are higher by an amount δ on average.

A similar argument can be made for patients who were excused from screening because they had already been screened at a clinic appointment in the preceding four weeks. We therefore included two additional covariates, Ref_i and Exc_i, in the model specified in section 7.6 as follows: For patients who had refused the offer of screening at any (or all) of 1, 2, 4 or 7 months, or at the auxiliary time point, we let the time-invariant indicator $Ref_i = 1$ (and $Ref_i = 0$ for patients who had never refused). Similarly, for patients who on at least one of those measurement occasions were excused from screening because they had been screened at a clinic appointment in the preceding four weeks, we let $Exc_i = 1$ (and $Exc_i = 0$ otherwise). Importantly, the addition of the two indicator variables had a large effect on predictions of missing scores from patients with no observed follow-up data because these effects when otherwise exercised through the joint modelling alone disappear when there are no observed scores to correlate with.

7.10 The refusal mechanism

Following on from the previous section, it seems possible that patients' *unobserved* distress scores were higher on those occasions when patients refused screening than on those occasions when screening was completed. The probability of a response being missing at time j due to patient refusal might therefore be dependent on the unobserved response at time j. The missing data process shown in section 7.6 is consequently no longer independent of the missing scores:

$$f(R_i \mid Y_i, \psi, X_i) = f(R_i \mid Y_i^O, Y_i^M, \psi, X_i) \neq f(R_i \mid Y_i^O, \psi, X_i)$$

The missing data process carries information about the unobserved values and can no longer be ignored when estimating θ or predicting Y^M. To proceed with the estimation task in the presence of a non-ignorable missing data likelihood it is therefore necessary to model the full joint likelihood for the responses and the missing data indicators:

$$f(Y_i \mid \theta, X_i)\, f(R_i \mid Y_i, \psi, X_i).$$

These functions are often intractable and typically no closed form can be derived, although it is possible to fit the likelihood function using numerical optimisation algorithms, for example as implemented in WinBUGS.

7.10.1 The missingness process

A pair of indicator variables was associated with patient i at time point j such that $\{R_{ij},M_{ij}\}=\{0,0\}$ for scores that were observed, $\{R_{ij},M_{ij}\}=\{1,0\}$ for scores that were missing due to refusals, and $\{R_{ij},M_{ij}\}=\{0,1\}$ for scores that were missing due to other reasons. The three events were modelled using two logistic regressions arranged in a hierarchical setup:

$$M_{ij} \sim Bernoulli(\pi_{Mij})$$

$$logit(\pi_{Mij}) = \alpha_{Mj}$$

and conditional on $M_{ij}=0$

$$R_{ij} \sim Bernoulli(\pi_{Rij})$$

$$logit(\pi_{Rij}) = \alpha_{Rj} + \beta_R y_{ij}$$

The regression for π_{Rij} modelled the conditional probability of an observation being missing due to refusal, conditional on the patient having been offered screening. Both regressions were parameterised with four separate intercepts to allow for varying levels of missingness at the four time points, although β_R, the effect of Y_{ij} on π_{Rij}, was assumed constant over time. α_{Rj} and α_{Mj} reflect the average levels of missingness due to refusals and other reasons respectively and were estimated directly from the data. What is the likely relationship between the value of the response and the probability that the patient refused screening? In the previous section we explored the relationship between patients' available responses and their propensity to have refused screening on at least one occasion. In a logistic regression we modelled the probability that patients refused screening on at least one occasion as a function of their average observed responses: Of patients with some available response data, around 7% had refused screening at least once. The estimated gradient of the regression (on the log odds scale) was 0.033 (95% CI: -0.005 to 0.071) suggesting that patients with average observed scores near the lower end of the HADS scale had a probability of around 4% of refusing screening one or more times. The equivalent probability near the top end of the HADS scale was around 15% (Table 7.7).

We assumed that the effect of Y_{ij} on π_{Rij} would be in the same direction but of a lesser magnitude (bearing in mind that the model was already controlling for the effect of being a Refuser). Instead of a single value we assigned a prior distribution to β_R since this approach allowed us to express the uncertainty about the parameter in a quantitative manner. We therefore specified the following prior distribution

$$\beta_R \sim N(0.0164, 0.0082^2)$$

which was consistent with an effect half the size, but recognising through the variance that it could be as big as 0.033 or it could be non-existent. As a sensitivity analysis, the model for the missing data indicators was fitted in WinBUGS jointly with the model for the incomplete response and covariate data.

Table 7.7. Modelled probabilities of patients refusing screening on at least one occasion as a function of their average observed distress score

Mean distress score	Modelled probability of refusal
0	4%
5	5%
10	6%
15	7%
20	8%
25	9%
30	11%
35	13%
40	15%

Based on a logistic regression with intercept: –3.08, and regression coefficient: 0.033.

7.11 Modelling informative missingness using offsets

A large proportion of patients had no observed response data during follow-up. Predictions of scores from these patients relied heavily on model assumptions as a consequence. Previously we have assumed that such patients were similar to the rest of the sample in terms of the levels of distress they experienced. However there was the possibility that many of them, particularly the ones who had had no appointments during follow-up, were in better health than the rest of the sample and consequently less distressed than patients with clinic appointments. Unless known to have been treated with palliative intent, or to have died within one year of follow-up, all patients without appointments during follow-up were assumed to belong to this group of healthier patients (in total 616 patients; 56% of all patients with no follow-up data).

As a further sensitivity model we therefore included a variable in the linear predictor of the response model to indicate whether patients satisfied the hypothesised criteria for being healthy and on long-term follow-up. We assessed the model results under a variety of values for this regression coefficient ranging from $\delta = 0$ equivalent to no departure from MAR, to $\delta = -2$.

7.12 Convergence

Having prepared the data for analysis using the SAS software we used a modified version of the publicly available SAS macro by Sparapani (2004) to transform the data into a format recognised by WinBUGS. The macro automatically assigns the value NA to any missing observations. To speed up convergence with the Gibbs sampler, the baseline HADS scores (range 15-42) were mean centred and rescaled to have unit variance to lessen the correlations between the intercepts and the regression coefficient and to speed up computation time.

Model parameter estimates were based on 5,000 simulated draws from the marginal posterior distributions of the parameters, having discarded the first 8000 iterations generated by the MCMC algorithm (the burn-in). We sampled from two chains simultaneously using over-dispersed initial values and calculating the BGR diagnostic (Brooks & Gelman, 1998) to assess convergence, coupled with visual inspection of the iteration histories.

7.13 Model estimates

The parameter estimates resulting from fitting the response model specified in sections 7.6 – 7.9 are shown in Table 7.8. There were only small differences between the intercepts during follow-up with a small decrease in scores over time. The strongest independent predictors of the responses were: time since diagnosis, deprivation score and baseline HADS score (range 15 to 42). There was some evidence that patients who had refused screening on at least one occasion scored a little higher on the HADS, whereas patients who had been excused from screening because they had already been screened at a clinic appointment in the preceding four weeks scored a little lower. The parameter estimates from the imputation model for the four incomplete covariates are shown in Table 7.9. The modelled covariates did not effect large variations in the responses although there were strong associations between these and other covariates, and amongst the incomplete covariates themselves.

Table 7.8. Estimated regression coefficients and variance parameters from fitting the response model.

	Posterior mean	(SD)
Intercepts		
1 month	14.32	1.69
2 months	13.96	1.67
4 months	13.78	1.67
7 months	13.50	1.66
Auxiliary time point	12.98	1.65
Gender		
male	0.36	0.63
Age		
50-64 years	0.80	0.46
≥65 years	0.68	0.47
Time since diagnosis		
1-5 years	1.26	0.40
>5 years	1.65	0.50
Cancer site		
breast	-0.19	0.96
genitourinary	0.46	1.22
gynae	-0.32	1.00
lower gastrointestinal	0.12	1.02
other	0.89	1.20
Treatment intent		
curative	0.00	0.39
Two months from start of treatment		
surgery	-0.21	0.54
radiotherapy	0.00	1.29
chemotherapy	-0.35	0.65
Health board		
Ayrshire and Arran	-1.14	1.57
Fife	1.88	1.52
Tayside	-0.82	1.52
Forth Valley	0.43	1.42
Lanarkshire	0.72	1.38
Argyll and Clyde	-0.21	1.37
Glasgow	0.97	1.33
Lothian	-0.11	1.34
SIMD quintile score		
2	-1.08	0.49
3	-1.09	0.52
4	-1.02	0.53
5	-1.36	0.50
Vital status 1 year post follow-up		
deceased	0.72	0.43
Standardised baseline HADS score	1.88	0.16
Screening refused at least once	0.87	0.54
Screening excused at least once	-0.70	0.42
Variance terms		
σ_e^2	15.49	0.72
σ_s^2	20.57	1.33
$\sigma_t^2 = \sigma_s^2 + \sigma_e^2$	36.06	1.27
$\rho = \sigma_s^2 / \sigma_t^2$	0.57	0.02

SIMD denotes Scottish Index of Multiple Deprivation. The reference categories were: female gender, <50 years of age, <1 year from diagnosis, benign cancer type, palliative treatment intent, health board: others, 1st SIMD quintile, alive one year post follow-up. Baseline HADS scores were standardised according to: standardised score = (baseline score – 18.75) / 3.70.

Table 7.9. Estimated regression coefficients (log odds scale) from fitting the models for the incomplete covariates.

	Curative treatment intent	Treatment started in the 2 months preceding baseline		
		Surgery	Radiotherapy	Chemotherapy
Intercept	2.32 (0.20)	0.42 (0.41)	-1.32 (0.60)	-1.64 (0.45)
Years since diagnosis	0.18 (0.04)	-6.71 (0.60)	-1.63 (0.41)	-3.28 (0.45)
Cancer site				
benign	-0.39 (0.45)	0.47 (0.55)	-51.24 (62.21)	-83.73 (60.77)
genitourinary	-2.26 (0.21)	-63.57 (61.34)	-0.59 (0.70)	-0.40 (0.72)
gynae	-0.91 (0.19)	-0.50 (0.28)	-63.50 (61.09)	1.53 (0.30)
lower gastrointestinal	-0.72 (0.20)	0.55 (0.25)	-0.27 (0.52)	0.44 (0.33)
other	-1.97 (0.24)	-1.43 (0.60)	0.07 (0.63)	0.27 (0.52)
SIMD quintile score				
2	-0.23 (0.20)	-0.76 (0.30)	-2.04 (0.87)	-0.04 (0.39)
3	-0.16 (0.21)	-0.31 (0.30)	-0.08 (0.55)	0.56 (0.38)
4	-0.19 (0.21)	-0.30 (0.30)	-0.92 (0.62)	0.45 (0.37)
5	0.16 (0.21)	-0.71 (0.28)	-1.06 (0.58)	0.45 (0.36)
Vital status 1 year post follow-up				
deceased	-1.64 (0.14)	-0.16 (0.34)	0.01 (0.58)	-0.30 (0.32)
Treatment intent				
curative	-	1.53 (0.33)	-0.52 (0.50)	-0.06 (0.31)
Treatment started in past two months				
surgery	0.76 (0.25)	-	-7.39 (12.82)	-0.90 (0.31)
radiotherapy	-0.64 (0.46)	-1.97 (0.94)	-	1.06 (0.50)
chemotherapy	-0.38 (0.29)	-0.93 (0.30)	0.98 (0.52)	-

Numbers are posterior means (standard deviations). SIMD denotes Scottish Index of Multiple Deprivation. The reference categories are: <1 year from diagnosis, breast cancer, 1st SIMD quintile, alive one year post follow-up.

The estimated intra-patient correlation between the repeated measurements was 0.57. This was somewhat lower than the first estimate derived in section 7.5 of around 0.64. However the estimates of the variance-covariance parameters presented in Table 7.8 are conditional on all covariates included in the model. The conditional, total variance of the responses, $Var(Y_{ij} \mid \alpha_j + X_i^T\beta) = \sigma_t^2 = \sigma_s^2 + \sigma_e^2$, is less because some of the variability that in section 7.5 was attributed to the inter-patient variance term, σ_s^2, is now explained by the added covariates. The error term is practically unchanged from that estimated in section 7.5 because the covariates are time-invariant; they explain variation between patients, not between time points.

7.14 Mean profiles and prevalence estimates

The average distress level in the sample was estimated at each of the four follow-up time points. This was done by pooling the marginal means of the posterior predictive

distribution for the missing responses with the means of the observed elements of Y at one, two, four and seven months post baseline. We also estimated the prevalence of significant psychological distress at each time point. This was done by pooling $Pr(\widetilde{Y}_{ij} \geq 15)$, the predictive probability that the missing responses were 15 or more, with the proportion of observed scores greater than or equal to 15 at each time point (Table 7.10). Also shown for comparison are the equivalent estimates from the POD Study.

Both mean distress and prevalence estimates were derived at each step of the model development described in sections 7.6 – 7.11. The results presented in Table 7.12 are from a selection of these models.

7.14.1 Comparison of mean distress and prevalence estimates

The average distress score as estimated with the screening data dropped considerably between baseline and one month followed by a slow decrease over the remainder of the follow-up period. The estimated prevalence was halved from baseline to one month but reduced only modestly thereafter. The prevalence estimate at seven months was 46.4% with a 95% confidence interval from 42.7 to 50.0% derived from the multiple imputed data. We found very similar patterns of change over time with each of the alternative scenarios, although the absolute values differed somewhat depending on the model (Table 7.12).

Compared with the POD Study results, analysis with the screening data resulted in larger estimates of the mean distress scores and distress prevalence rates, but the patterns of change during follow-up in both mean scores and prevalence estimates were similar. There were many potential reasons for the level of distress being higher when basing the estimates on the screening data. Because of the large amount of missing data, the results obtained with the screening data depended largely on modelled data. The results were therefore highly sensitive to modelling assumptions, including assumptions about the structure of the relationship between the response variable and the covariate data, the validity and type of covariates included, unmeasured confounders, and the covariance structure between the repeated

responses. However there was also the possibility that the analysis with the screening data did not overestimate the parameters of interest. The discrepancies in the distress level might have reflected actual differences in the populations being studied. The cases included from the screening data were drawn from a geographically much wider area than those in the clinical study. There were more men in the screening sample (19% versus 14%, p=0.031), the screening sample was slightly older on average (62.3 versus 61.0 years, p=0.072) and bowel cancer was more than twice as prevalent than in the clinical study (17% versus 8%, p<0.001).

Perhaps the sub-sample of patients in the screening data on whom we had observed data during follow-up was not representative of the whole cohort. Perhaps people seen in hospital were more distressed on average than those not seen in hospital. We addressed this problem in part by allowing for a lower intercept in patients on long-term follow-up who were not seen in hospital, and assessed the findings under a range of plausible values for the intercept offset.

The POD Study collected follow-up data from patients at home over the telephone, but the screening data were collected from patients who were waiting for appointments in hospital clinics. Perhaps the distress scores collected in the clinic area were transiently inflated due to the patient's anticipation of the upcoming appointment, and the hospital surroundings. This would be similar to the so-called 'white-coat' effect where patients exhibit higher blood pressure levels when measured in a clinical setting than when measured at home. We investigated this possibility in detail in Chapter 4 and found some evidence that patients scored around 0.5 units higher on the HADS when measured in the clinic instead of at home over the telephone. We also found that patients who refused screening were more distressed on average. The analysis of the screening data included all eligible cases initially identified with distress, but the POD Study only followed up patients who had agreed to take part in a clinical study. It is likely that this self-selected group consisted of fewer 'refusers' than did the screening data cohort.

7.15 Multiple imputed datasets

The second aim of the study was to determine whether early predictors of long-term distress existed where long-term distress was defined as significant distress at seven months, $Y_{i4} \geq 15$.

The POD Study addressed this research question using a multivariable logistic regression to estimate associations with a set of demographic and disease characteristics. We wanted to repeat this analysis with the routinely collected screening data. However a logistic regression including only patients with complete covariate data and available data at one and seven months would limit the analysis to include only a fraction of the original sample.

Fortunately the response model that was developed in the previous sections could be used to predict the missing response and covariate data. We generated multiple imputed datasets in WinBUGS by drawing from the posterior predictive distribution for the missing response and covariate data. Consecutive draws obtained using the Gibbs sampler are not ordinarily independent. We therefore retained only every 100th iteration of the sampler to ensure a lag between the predictions large enough that no autocorrelation remained in the chain. Using this method we extracted 100 independent sets of imputations. The imputed data were saved in data files and imported into SAS where they were merged with the observed data to create a sequence of 100 completed datasets. For each completed dataset the dichotomised distress scores at seven months were derived from the continuous Y variables and analysed in a logistic regression. Finally the resultant 100 sets of estimates were combined into a single set of estimates based on multiple imputation theory. We carried out this step in SAS and used the MI procedure to combine the multiple sets of estimates.

7.16 Associations with distress at seven months

The aim as set out in the POD Study was to determine the demographic, cancer and distress characteristics at baseline that were predictive of persistent distress defined as a HADS score ≥ 15 at seven months. The study included the following covariates

in the logistic regression: gender, age (<50 years/ 50 – 64 years/ ≥65 years), cancer site (bowel/ breast/ genitourinary/ gynaecological/ other), disease activity (disease-free/ active disease), cancer treatments received in the two months preceding baseline (no treatment/ chemo or radiotherapy/ surgery only/ hormone treatment only), marital status (not married/ married) and baseline distress score (HADS<20/ HADS≥20). These were the variables available at baseline. In a subsequent analysis, distress status at one month was added to the list of covariates as it was hypothesised that a second high distress score one month after the baseline screening would predict distress six months later.

Details on disease activity and marital status were not available in the screening data, and these covariates were consequently left out of the present predictor analysis. The screening data also only contained details on treatment start dates, but not the duration of treatments or dates of completion. The covariates from the screening data therefore only indicated whether treatments were started in the two months preceding baseline which is subtly different from whether treatments were received in that same period. For the purpose of comparing results from analysis of the POD Study with results from the screening data, we repeated the POD Study predictor analysis, but leaving out marital status and disease activity. The focus in the present chapter will be on the fully inclusive multivariable analysis which is adjusted for baseline covariates as well as distress status at one month. Results from the POD Study are compared with those from the screening data in Table 7.11.

Table 7.10. Comparison of mean levels and prevalence of distress during follow-up estimated from the screening data and POD Study data.

	Screening data		POD Study data	
	N §	Posterior mean (SD)	N §	Mean (SE)
HADS score (range 0–42)				
Baseline	2149+ 0	18.8 (0.08) #	325+ 0	18.7 (0.20)
1 month	226+1923	15.2 (0.34)	325+ 0	13.4 (0.31)
2 months	437+1712	14.9 (0.25)	307+18	13.7 (0.35)
4 months	499+1650	14.7 (0.22)	296+29	13.6 (0.37)
7 months	490+1659	14.4 (0.23)	292+33	12.7 (0.39)
Distress prevalence				
Baseline	2149+ 0	100%	325+ 0	100%
1 month	226+1923	51.4% (2.5%)	325+ 0	40.0% (2.7%)
2 months	437+1712	49.7% (2.0%)	307+18	43.2% (2.8%)
4 months	499+1650	48.4% (1.7%)	296+29	41.6% (2.8%)
7 months	490+1659	46.4% (1.8%)	292+33	37.8% (2.8%)

Distress defined as HADS score ≥ 15. Distress prevalence at baseline was 100% by design. # The distribution for baseline HADS score was not estimated in the Bayesian analysis since this variable entered the model as a fully observed covariate. The reported numbers are the sample mean and (frequentist) standard error. Missing data were handled using multiple imputation.§ Numbers are scores observed + scores imputed.

Table 7.11. Comparison of estimated associations of persistent distress based on analysis with the screening data and the POD Study data *(continued on next page)*.

	Screening data	POD Study data
	Adjusted log odds ratio (standard error)	**Adjusted log odds ratio (standard error)**
Total		
Gender		
Female	0	0
Male	0.24 (0.26)	0.28 (0.81)
Age (years)		
< 50	0	0
50 to 64	0.23 (0.20)	-0.01 (0.35)
≥ 65	0.19 (0.20)	-0.21 (0.37)
Primary cancer		
Bowel	0	0
Breast	0.10 (0.23)	-0.36 (0.68)
Genitourinary	0.13 (0.33)	-0.27 (0.78)
Gynaecological	0.02 (0.25)	-0.32 (0.71)
Other	0.13 (0.28)	-0.85 (0.84)
Cancer treatment	**§**	**#**
Surgery/chemo or radiotherapy	0	0
None or hormone therapy only	0.16 (0.18)	0.70 (0.29)
Baseline HADS		
Score < 20	0	0
Score ≥ 20	0.62 (0.14)	0.26 (0.29)
Distress status at 1 month		
No significant distress	0	0
Significant distress	1.80 (0.16)	1.70 (0.28)

Table 7.11. *(Continued)*.

	Screening data		POD Study data	
	Adjusted odds ratio (95% CI)	**P-value**	**Adjusted odds ratio (95% CI)**	**P-value**
Total				
Gender		**0.353**		**0.725**
Female	1		1	
Male	1.27 (0.77 to 2.09)		1.33 (0.27 to 6.48)	
Age (years)		**0.509**		**0.761**
< 50	1		1	
50 to 64	1.25 (0.85 to 1.85)		0.99 (0.50 to 1.99)	
≥ 65	1.21 (0.82 to 1.79)		0.81 (0.39 to 1.67)	
Primary cancer		**0.974**		**0.906**
Bowel	1		1	
Breast	1.11 (0.70 to 1.75)		0.70 (0.19 to 2.63)	
Genitourinary	1.14 (0.60 to 2.16)		0.76 (0.17 to 3.51)	
Gynaecological	1.02 (0.63 to 1.66)		0.73 (0.18 to 2.95)	
Other	1.13 (0.65 to 1.99)		0.43 (0.08 to 2.21)	
Cancer treatment	**§**	**0.373**	**#**	**0.015**
Surgery/chemo or radiotherapy	1		1	
None or hormone therapy only	1.17 (0.83 to 1.65)		2.01 (1.15 to 3.54)	
Baseline HADS		**<.001**		**0.361**
Score < 20	1		1	
Score ≥ 20	1.87 (1.41 to 2.47)		1.30 (0.74 to 2.29)	
Distress status at 1 month		**<.001**		**<.001**
No significant distress	1		1	
Significant distress	6.03 (4.37 to 8.32)		5.45 (3.15 to 9.42)	

Parameter estimates are conditional on all other variables in the model. Missing data were handled using multiple imputation. § Treatments *started* in the two months preceding baseline. # Treatments *received* in the two months preceding baseline.

7.16.1 Comparison of results from the predictor analysis

The two analyses resulting from the screening data and the POD Study data were in agreement that there were no significant associations of persistent distress at seven months with gender, age and cancer site, and that a strong association existed with distress status at one month. The POD Study results suggested that patients not in receipt of cancer treatments at baseline were more likely to remain distressed at seven months (OR: 2.01, 95% CI: 1.15 to 3.54). The point estimate of the equivalent effect in the screening data was smaller and not statistically significant (OR: 1.17, p=0.373) but did not contradict the finding and was contained within the 95% confidence interval for the estimated effect in the POD Study. The analysis with the

screening data suggested that the baseline HADS score was a significant predictor, but only when distress at one month was left out of the multivariable model did the POD Study confirm this finding (OR: 1.89, 95% CI: 1.13 to 3.16; results not shown in Table 7.11).

Again there were several possible reasons for these discrepancies. Of the four follow-up time points, the scores at one month were the least observed in the screening data. Estimated associations involving distress status at one month therefore relied heavily on model predictions of scores at one month. These in turn were determined by the dependencies set out in the model. Via the random intercept the model imposed a dependency between the scores at one and seven months while simultaneously stipulating a constant effect of the baseline score on the scores at each of one, two, four and seven months. Hence, it is not surprising that the modelled screening data suggest a statistically significant effect of the baseline distress variable independent of distress at one month. The independent effect of having a high baseline distress score was estimated in the POD Study to increase the odds of distress at seven months by a factor of 1.30 (95% CI, 0.74 to 2.29). The confidence interval for this effect comfortably includes 1.87 which is the point estimate of the same effect derived with the screening data. There was no evidence therefore that the results were contradictory. Lastly, we noted above that the variables available on the cancer treatments received by patients in the POD Study and in the screening data were not entirely equivalent. This could have explained some of the difference in the estimated associations with persistent distress at seven months. In summary, the results from the two analyses were rather similar and with no directly contradictory results.

The predictor analysis was repeated with data imputed under the models of section 7.10 and 7.11. Results from analysis under these alternative models were very similar to those under the main model (Table 7.13).

Table 7.12. Mean levels and prevalence of distress during follow-up estimated from the screening data under alternative model assumptions.

	Informative refusal mechanism, β_R=*0.0164*	MNAR offset for patients on long-term follow-up δ= -0.5	δ= -1.0	δ= -2.0
HADS score (range 0–42)				
Baseline	18.8 (0.08)#	18.8 (0.08)#	18.8 (0.08)#	18.8 (0.08)#
1 month	15.7 (0.35)	15.1 (0.34)	15.0 (0.34)	14.9 (0.34)
2 months	15.3 (0.25)	14.8 (0.25)	14.7 (0.25)	14.5 (0.25)
4 months	15.1 (0.23)	14.6 (0.22)	14.5 (0.22)	14.3 (0.22)
7 months	15.0 (0.24)	14.3 (0.23)	14.2 (0.23)	14.0 (0.23)
Distress prevalence				
Baseline	100%	100%	100%	100%
1 month	53.8% (2.4%)	50.8% (2.5%)	50.2% (2.5%)	49.0% (2.5%)
2 months	51.9% (1.8%)	49.1% (1.9%)	48.5% (1.9%)	47.3% (1.9%)
4 months	50.7% (1.7%)	47.7% (1.7%)	47.1% (1.7%)	45.8% (1.7%)
7 months	49.5% (1.7%)	45.7% (1.8%)	45.1% (1.7%)	43.8% (1.8%)

Data are posterior means (SD). Distress defined as HADS score ≥ 15. Distress prevalence at baseline was 100% by design. # The distribution for baseline HADS score was not estimated in the Bayesian analysis since this variable entered the model as a fully observed covariate. The reported numbers are the sample mean and (frequentist) standard error.

Table 7.13. Analysis of the screening data imputed under the alternative models of sections 7.10 – 7.11. Prevalence of distress at seven months and associations with patient characteristics *(Continued on next page)*.

	Multivariate analysis			
	Non-ignorable refusal mechanism, β_R=0.0164.		MNAR missingness through δ =-2 offset for patients on long-term follow-up.	
	Odds ratio *(95% CI)*	**P-value**	**Odds ratio *(95% CI)***	**P-value**
Total				
Gender		**0.463**		**0.290**
Female	1		1	
Male	1.21 (0.73 to 2.02)		1.32 (0.79 to 2.19)	
Age (years)		**0.527**		**0.535**
< 50	1		1	
50 to 64	1.24 (0.85 to 1.82)		1.25 (0.84 to 1.86)	
≥ 65	1.23 (0.83 to 1.83)		1.20 (0.80 to 1.81)	
Primary cancer		**0.951**		**0.993**
Bowel	1		1	
Breast	1.03 (0.68 to 1.58)		1.08 (0.69 to 1.69)	
Genitourinary	1.26 (0.69 to 2.33)		1.13 (0.60 to 2.14)	
Gynaecological	0.98 (0.60 to 1.59)		1.07 (0.65 to 1.75)	
Other	1.08 (0.63 to 1.87)		1.10 (0.62 to 1.95)	
Cancer treatment ¥		**0.387**		**0.567**
Surgery/chemo or radiotherapy	1		1	
None or hormone therapy only	1.16 (0.83 to 1.64)		1.10 (0.79 to 1.55)	
Baseline HADS		**<.001**		**<.001**
Score < 20	1		1	
Score ≥ 20	1.83 (1.38 to 2.42)		1.86 (1.40 to 2.46)	
Distress status at 1 month		**<.001**		**<.001**
No significant distress	1		1	
Significant distress	6.39 (4.72 to 8.66)		6.24 (4.51 to 8.65)	

Table 7.14. *(Continued)*.

	Multivariate analysis	
	Non-ignorable refusal mechanism, β_R=0.0164, and δ =-2 offset for patients on long-term follow-up.	
	Odds ratio *(95% CI)*	**P-value**
Total		
Gender		**0.413**
Female	1	
Male	1.25 (0.73 to 2.14)	
Age (years)		**0.572**
< 50	1	
50 to 64	1.23 (0.83 to 1.81)	
≥ 65	1.22 (0.82 to 1.81)	
Primary cancer		**0.980**
Bowel	1	
Breast	1.01 (0.64 to 1.58)	
Genitourinary	1.22 (0.65 to 2.28)	
Gynaecological	1.02 (0.62 to 1.68)	
Other	1.06 (0.62 to 1.81)	
Cancer treatment ¥		**0.500**
Surgery/chemo or radiotherapy	1	
None or hormone therapy only	1.12 (0.80 to 1.58)	
Baseline HADS		**<.001**
Score < 20	1	
Score ≥ 20	1.82 (1.38 to 2.39)	
Distress status at 1 month		**<.001**
No significant distress	1	
Significant distress	6.57 (4.88 to 8.85)	

¥Treatments started in the two months preceding baseline.

7.17 Discussion

7.17.1 Limitations

A main focus throughout this chapter has been the limitations of the screening data and the methodological approaches used to analyse these. To provide a repeat discussion of these in this section serves little purpose, however the following few points have not been made previously.

The linked oncology data provided details related to each primary cancer, but did not include information on recurrences. When patients had more than one primary cancer recorded, we used information pertaining to the most recent cancer prior to their baseline distress measurement. The nature of the clinical data was such that we could not reliably assess the effect of time-varying covariates, such as starting treatment, or having a second primary cancer diagnosis during follow-up. We were also unable to link reasons for patients' clinic visits, or the content of their consultations, to variations in distress profiles because the screening data did not include details about the consultations. Furthermore, not all sections of the hospital oncology clinics were monitored. There is therefore a possibility that some patients may have had clinic appointments that were not recorded in the screening database.

We also found that there were too few observations to estimate reliably the covariate effects separately at each time point. The model was therefore restricted to include only constant effects of the covariates over time. This was likely to be an oversimplification. Particularly the treatment variables and the baseline distress score may have had relatively larger effects early on during follow-up. Ideally, with more observed data, the model could have accommodated time interactions with some of the covariates. Having selected cases for analysis from the irregularly spaced screening data we forced the measurements into a regular framework to match the structure of the PODS dataset. As a result some data were discarded where multiple measurements existed within a time window. As an alternative to this, a more efficient approach might have applied an analysis with the screening data that was capable of utilising the data in its original irregular form. Finally, the data included in the analysis from the nominal time points at one, two, four and seven months could

have been selected differently, and different estimates might have resulted as a consequence. Also, the functional form of the imputation model may not have been representative of the true dependencies in the data.

The choice of 'non-informative' prior distribution for the subject-level variance parameters in longitudinal models is not straightforward. We specified an inverse-gamma prior distribution for σ_s^2. Although this choice of prior has been suggested in the past with hierarchical models (Spiegelhalter et al., 2003), Gelman (2006) advises against the use of the inverse-gamma prior distribution on variance parameters in hierarchical models when the aim is to specify a weak prior relative to the data likelihood. The article uses an example to illustrate how the choice of 'non-informative' prior can have a big effect on the posterior distribution. The problem is more serious when the group-level variance is close to zero and if the number of groups (subjects in our example) is small because the data then contain little information about the group-level variance. The article recommends using a non-informative uniform prior or a half-normal centred at zero for the class-level standard deviation in place of the inverse-gamma distribution on the variance. We refitted the main model from this chapter using as a prior for σ_s the half-normal distribution centred at zero with variance equal to 100^2 but found no discernible change in the posterior for σ_s or indeed for any of the other model parameters, perhaps because of the large number of subjects and relatively large subject-level variance.

7.17.2 In conclusion

In this chapter we presented an analysis of routinely collected psychological screening data from oncology outpatient clinics. We aimed to determine whether analysis of such incomplete data could be used to address research questions by comparing the results with those of a purpose-designed clinical study. Overall we obtained similar results with the two analyses. We also found that the average level of distress and the prevalence of distress cases were somewhat higher in the screening data. We gave a number of possible reasons why the results from the two analyses might not agree entirely. It was not possible to conclude definitively whether the POD Study population was somewhat less distressed or if the prevalence

estimates were different as a result of bias in the screening data. The sensitivity models yielded largely similar results to the main model; perhaps the model assumptions behind these were not extreme enough to capture the actual confounding mechanisms at play.

Both the findings from the predictor analysis and the estimated regression coefficients of the imputation model suggested that only very few of the covariates were associated with the response variable and that none of the associations were particularly strong. This is important because the analysis relied on dependencies within the observed data to model the incomplete responses. The lack of strong associations may have been a consequence both of the design, whereby only patients with a high baseline score were followed up resulting in a strong regression-to-mean effect, and the choice of a psychological self-report as the response variable, which is known to have a large random error. Given the large amount of missingness in the screening data it is not surprising then that it proved a challenge to reproduce exactly the results obtained in the POD Study.

8 FURTHER ANALYSIS WITH THE SCREENING DATA

In this chapter we will again consider the routine data collected by the screening service, but this time for original research purposes addressing questions that have not been adequately studied elsewhere. The chapter will focus on the development of methods for an analysis of incomplete two-stage screening data.

8.1 Background

Depression is thought to be a relatively common problem among people who have had a cancer diagnosis although exactly how common is unclear. A recent systematic review (Walker et al., 2012) found that published prevalence estimates vary wildly due to small and poorly executed studies. Consequently there is a need for a good, large-scale study of the prevalence of depression in clinically meaningful subgroups of people who have had a cancer diagnosis.

The Symptom Monitoring Service routinely screened patients for symptoms of depression when they attended appointments in the screened cancer outpatient clinics. The service screened more than 20,000 patients using diagnostic interviews to assess patients for major depression. There are no other studies of depression in a cancer outpatient population that are based on sample sizes of this magnitude. Based on analysis with the screening data we therefore aimed to:

(1) estimate the prevalence of depression among patients with a cancer diagnosis attending outpatient oncology clinics

(2) identify patient demographic and clinical characteristics associated with major depression

8.2 The data

The Symptom Monitoring Service (further details in Chapters 4, 5 and 7) was a depression screening service that operated in a large number of NHS cancer outpatient clinics in central Scotland covering a geographically defined area of about four million people. In advance of each clinic the screening service received a list

with the details of patients scheduled to attend for appointment. When attending the clinics, patients were approached by the screening service and asked to complete a brief questionnaire asking them about physical and psychological symptoms common to cancer patients. The questionnaire included the Hospital Anxiety and Depression Scale (HADS), the details of which have been described in previous chapters. The HADS total scale ranges from zero (no distress) to 42 (maximal distress), and a HADS total score of 15 or more has been shown to be a good indicator of significant psychological distress warranting further action when screening for depression. Patients who scored 15 or more on the HADS in clinic were subsequently telephoned at home and asked to complete an interview for depression using the major depression component of the Structured Clinical Interview for DSM-IV (SCID; First et al., 1999). Patients diagnosed with major depression were also asked about any antidepressant medication or psychological treatments they were receiving for their depression.

In addition to the screening data we also obtained linked demographic and clinical records from the Scottish Cancer Registry (see Chapter 7) on almost all patients in the screening database. Over the period during which the screening service operated more than 30,000 people were scheduled for appointments in the screened oncology outpatient clinics. Over 24,000 patients completed the screening questionnaires and most (97%) agreed for their questionnaire data and clinical data to be used for research purposes (in anonymised and aggregated form) and were successfully matched to their clinical oncology data held by the Scottish Cancer Registry (Figure 8.1). The final linked research database was prepared in anonymised form by NHS Scotland Information Services Division (ISD).

8.3 A layered approach to the analysis

The screening service operated a two-stage screening service whereby patients were screened first using the HADS, and second, if they had scored high on the HADS, over the telephone using a clinical interview for depression. The idea behind this procedure was that in principle only patients with a high clinic HADS score can be depressed, and to interview all patients (without screening for high HADS scores)

would therefore be wasteful. (We will return to this assumption later in section 8.11). Yet even of those who had scored high on the HADS, 21% (1276/6108) did not complete the clinical interview for depression.

The prevalence estimation problem can be thought of as existing across a number of layers. The inner-most layer consists of patients who either scored low on the HADS in clinic (score<15), or scored high on the HADS and completed the subsequent depression interview. Within this layer we can be very confident about the data, but a prevalence estimate based on this subgroup has little external validity.

The second layer consists of all patients who completed the HADS in clinic, some of whom scored high but failed to complete the subsequent depression interview. While it is clear that the depression diagnosis is not MCAR in this case, it may be reasonable to assume that it is MAR conditional on HADS≥15. Within this layer we have to make some assumptions about the unobserved data but there is greater external validity.

The final layer consists of all patients with at least one scheduled appointment in one of the screened clinics. This layer includes patients who completed screening but did not want their data used for research, patients who were not matched to their clinical records, and nearly 7,000 patients who never completed screening. Clearly, an estimate based on such a sample would have good external validity, although one would have to make strong assumptions about the unobserved data.

In the following we will approach the analysis within this framework, starting with the data we are most confident about.

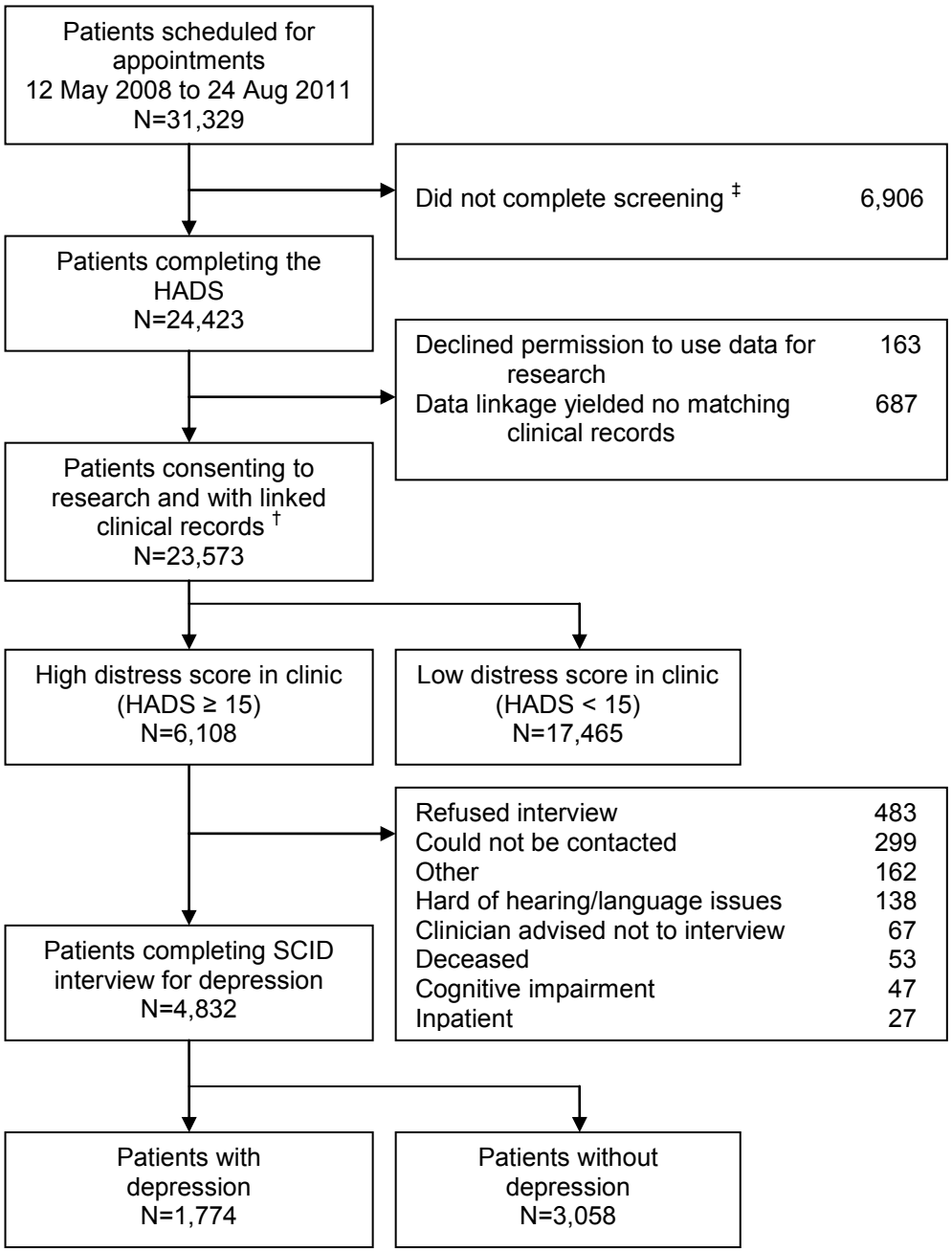


Figure 8.1. Derivation of the analysis sample. ‡ Reasons included patients not attending appointments, patients missed in the clinic, patients declining screening, clinicians advising not to screen, patients too unwell or unable to use screening. † If patients made multiple clinic visits, the first visit on which the HADS was available was used in the analysis.

8.4 Choosing a single observation from each patient

More than 24,000 patients completed the HADS questionnaire, but many patients did so more than once. Put together, patients completed the questionnaire more than 50,000 times. When patients completed the questionnaire more than once there was a choice therefore as to which questionnaire to include in the analysis. We considered three alternatives.

First, the questionnaire to be included in the analysis could be chosen at random from the set of available questionnaires from each patient. (This rule could also be based on appointments rather than available questionnaires, such that the appointment included for analysis is chosen at random from the patient's history of appointments. Missing data arising from appointments with no available questionnaire scores would then have to be handled in the analysis.) One potential problem with this approach is that many patients came to clinic only once; consequently many of the scores would originate from the patients' first clinic visit.

More straightforward perhaps is the second option whereby the patient's first appointment is always selected for inclusion in the analysis and any subsequent appointments are ignored. Again with this approach the sample would be subject to a substantial degree of missingness.

The third option is more pragmatic and would base the analysis on the first available questionnaire from each patient. Therefore if a patient attended the clinics twice, but only completed the questionnaire on the second occasion (or declined research consent on the first occasion), the analysis would be based on data collected at the second appointment.

The main analysis presented in this chapter was based on option three above. Figure 8.1 shows the derivation of the analysis sample. There was a potential risk of bias associated with this option because patients' psychological distress levels might have differed systematically on those occasions when they did not complete the screening,

although major depression is more persistent in nature than other types of psychological distress.

8.5 The problem with an overall estimate

A first overall estimate of the depression prevalence was based on the 23,573 patients with available HADS scores. Of these, 6,108 (26%) scored high (score≥15) on the HADS and were shortlisted for further investigation, 4,832 patients actually completed the depression interview, and 1,774 (37%) of these were found to have major depression. So how can we derive an overall prevalence estimate from these figures?

Of the 23,573 patients who completed the HADS, 1,276 patients (6,108 – 4,832) had a missing depression diagnosis, all of whom had scored high on the HADS. A naïve analysis might derive the number of confirmed depression cases as a proportion of the total number of patients, 1774/23573=7.5%. This produces an obvious underestimate of the true rate of depression among patients who completed the HADS. The analysis assumes that none of the 1,276 patients with a high HADS score and a missing depression diagnosis were actually depressed. A more likely assumption was that the rate of depression among this group of patients was the same as that in the group who did complete the interview. (Assuming that the depression diagnosis is missing at random conditional on the clinic HADS score.)

Hence, the overall depression estimate is the product of the rate of depression among patients with HADS≥15 *and* the proportion of patients with a high HADS score:

$$\hat{p} = \frac{1774}{4832} x \frac{6108}{23573} = 0.095.$$

So, an overall estimate of the prevalence of depression among cancer outpatients attending oncology clinics who complete the screening questionnaire in clinic is 9.5%.

However it is difficult to know how this estimate generalises to an external, meaningful population, the reason being that the depression screening service targeted clinics that specialised in cancer types carrying a high risk of major depression. For example, the screening service operated in Urology clinics for a limited period only because major depression proved not to be very prevalent among patients who visited these clinics; as a consequence, to optimise the use of resources, the screening service chose to focus on other cancer types. The estimate is therefore biased due to an overrepresentation of patients with certain cancer types known to be associated with major depression.

Aside from this issue it is also questionable to what extent an overall estimate across all cancers is clinically useful. More relevant to clinical practice perhaps are prevalence estimates among subgroups of patients, for example as defined by the cancer site.

8.6 The patient data

The most common cancer types represented in the screening database were breast cancer (N=8,462), lung cancer (N=4,316), lower gastrointestinal cancers (N=3,356), gynaecological cancers (N=3,010), and genitourinary cancer (N=2,009). As these are the five most common cancer types the analysis was limited to focus on these.

The patient characteristics collected by the screening service and from the linked clinical records are presented in Tables 8.1 and 8.2. Complete data were available on most variables. The notable exceptions were clinic appointment type (1% missing), the treatment variables (4%, 3% and 6% missing for chemotherapy, radiotherapy and surgery respectively), therapeutic objective (12% missing) and prognosis for patients with a high clinic HADS score (9% missing). The screening service helped identify patients that were eligible for clinical trial participation, and for that purpose the service enquired about prognoses for patients with a high HADS score.

There were 46 patients whose date of diagnosis as registered by the Scottish Cancer Registry was after their first appointments at the screened oncology clinics. This may

highlight inaccuracies in the cancer registry; however out of more than 20,000 registrations, this level of inconsistency is hardly alarming. The registered diagnosis date for most of the 46 patients was very soon after the clinic appointment. For six patients however, the diagnosis date was between one and two years after the first clinic appointment perhaps suggesting that the cancer registered in the database might not have been the first primary cancer for those patients.

In the analysis that follows we will estimate the prevalence of depression separately within each of the five main cancer types listed above. We will also analyse the data for associations between patient characteristics and major depressive disorder as this may help clinicians identify cancer patients at increased risk of depression.

Table 8.1. Patients' demographic and clinic characteristics.

Total	21,153 (100%)
Gender	
Female	15113 (71%)
Male	6040 (29%)
Age (years)	
mean (SD) min – max	64.4 (11.9) 19 – 100
median (IQR)	65 (57 – 73)
Age group	
< 50 years	2521 (12%)
50 to 59 years	4105 (19%)
60 to 69 years	6820 (32%)
≥ 70 years	7707 (36%)
Health board*	
Lothian	6287 (30%)
Glasgow	4320 (20%)
Argyll & Clyde	2665 (13%)
Lanarkshire	2558 (12%)
Forth Valley	1615 (8%)
Fife	1246 (6%)
Tayside	1592 (8%)
Other†	868 (4%)
Resident setting	
Urban	16689 (79%)
Small town	2001 (9%)
Rural	2461 (12%)
Deprivation SIMD quintile score**	
1	4572 (22%)
2	4259 (20%)
3	3781 (18%)
4	3731 (18%)
5	4808 (23%)
Clinic appointment type	
First appointment	3118 (15%)
Return appointment	17761 (84%)
Missing data	274 (1%)
Clinic HADS score	
mean (SD) min – max	10.4 (7.5) 0 – 42
median (IQR)	9 (5 – 15)
Clinic distress status	
HADS<15	15641 (74%)
HADS≥15	5512 (26%)

Data are number (%) unless otherwise indicated. * Health board (HB) where patient was resident when cancer was registered. † Ayrshire and Arran, Borders, Dumfries & Galloway, Grampian and Western Isles. ** Scottish Index of Multiple Deprivation quintile score: 1=most deprived, 5=least deprived. Note: in addition to other missing data indicated, health board, resident setting and deprivation score were missing for two patients.

Table 8.2. Patients' disease and treatment characteristics.

Total	21153 (100%)
Cancer type	
Lower gastrointestinal	3356 (16%)
Breast	8462 (40%)
Genitourinary	2009 (9%)
Gynaecological	3010 (14%)
Lung	4316 (20%)
Time since diagnosis (years)	
mean (SD) min; max	2.4 (3.3) -2.0 – 31.5
median (IQR)	1.0 (0.3 – 3.2)
Chemotherapy treatment started	
Yes	11028 (52%)
No	9331 (44%)
Missing data	794 (4%)
If yes, time since treatment started (months)	
mean (SD) min; max	1.8 (2.5) -1.7 – 14.0
median (IQR)	0.7 (0.1 – 2.8)
Radiotherapy treatment started	
Yes	11070 (52%)
No	8764 (41%)
Missing data	1319 (6%)
If yes, time since treatment started (months)	
mean (SD) min; max	1.7 (2.7) -1.7– 13.3
median (IQR)	0.4 (0 – 2.8)
Surgical treatment received	
Yes	14218 (67%)
No	6365 (30%)
Missing data	570 (3%)
If yes, time since surgery (months)	
mean (SD) min; max	2.4 (2.8) -2.3 – 14.1
median (IQR)	1.2 (0.3 – 3.7)
Therapy objective	
Curative	13128 (62%)
Palliative	5413 (26%)
Missing data	2612 (12%)
Poor prognosis (patients with HADS≥15 only)#	
Yes	521 (9%)
No	4469 (81%)
Missing data	522 (9%)

Data are number (%) unless otherwise indicated. # Poor prognosis was defined for lung (non-lung) cancer patients as a life expectancy of less than three (twelve) months. For purposes of clinical trial recruitment the screening service enquired about prognoses for patients with a high HADS score.

8.7 The response model

The two patients with missing data on health board, resident setting and deprivation score were removed from the analysis. However, the incompletely observed variables on the remaining 21,151 patients were handled in the analysis.

About one fifth of patients who had scored high on the distress measure in clinic failed to complete the clinical interview for depression, and the depression status (depressed/not depressed) was consequently missing for these patients. (In section 8.11 we shall revisit the assumption that only patients with a high distress score can be depressed).

Conditional on covariate data it seems reasonable to assume a similar level of depression among these patients as was observed among patients who completed the depression interview. This corresponds to a MAR assumption conditional on covariate data. Therefore in predicting the missing depression diagnoses we modelled the dependencies between the covariate data and the response variable, depression (yes/no), on the dependencies observed among the high scorers (HADS≥15) who had completed the depression interview.

Using Bayesian methods in WinBUGS we fitted the following model to the data observed from patients with a clinic HADS of 15 or more.

$$Y_i \sim Bernoulli(\pi_i) \qquad \text{(model 8.1)}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{X}_i^T\boldsymbol{\beta}$$

Through the log odds link function the logistic regression models the probability π_i that patient i was depressed ($Y_i=1$). The probability is expressed as a function of the covariate terms $X_i^T\beta$ where X_i^T is a row vector of covariate data belonging to patient i and β is the vector of corresponding regression coefficients. We included most of the covariate data listed in Tables 8.1 and 8.2 in the linear predictor as the primary

purpose of this model was to impute the missing depression diagnoses. Thus, we regressed the outcome on the following variables: gender, age group, health board, resident setting, deprivation score, clinic HADS score, cancer type, time since diagnosis (<3 months, 3-12 months, 1-3 years, >3 years), recently started on treatment (received surgery or started chemotherapy or radiotherapy in preceding six months) and therapy objective. We specified a multivariate normal distribution for $\boldsymbol{\beta}$ with mean vector $\boldsymbol{\mu}=\boldsymbol{0}$, and a diagonal precision matrix $\boldsymbol{\Sigma}^{-1}$ with all entries equal to *0.001*.

8.8 The covariate model

The relevant four incompletely observed covariates were: therapeutic objective [curative/palliative] and the three binary treatment variables, chemotherapy, radiotherapy and surgery started in the preceding six months. The variables were modelled jointly in a set of linked logistic regressions as in Chapter 7.

Consider the vector of covariates, $\boldsymbol{X}_i$, as consisting of two parts. The first part, $\boldsymbol{W}$, consists of the completely observed elements of $\boldsymbol{X}_i$ and are thought of as non-stochastic constants while the second part, $\boldsymbol{Q}_i$, consists of the incompletely observed elements of $\boldsymbol{X}_i$. The elements of $\boldsymbol{Q}_i$ are treated in the analysis as random variables.

We also define $\boldsymbol{W}_i'$ to be a subset of $\boldsymbol{W}_i$ that consists of variables that are predictive of the missing $\boldsymbol{Q}_i$. Through exploratory analyses we found the following completely observed covariates to be relevant in predicting the four incomplete covariates: health board, age category, deprivation score, cancer type and time since diagnosis (<3 months, 3-12 months, 1-3 years, >3 years).

Letting $\boldsymbol{Q}_{ik}$ denote the *k*th covariate of the $\boldsymbol{Q}$ vector from patient *i* we specified the joint distribution for the four incomplete binary covariates $Q_1,\ldots, Q_4$ through a series of chained logistic regressions:

$$Q_{ik} \sim Bernoulli(\pi_{ik}) \quad , \qquad k=1,\ldots,4$$

(model 8.2)

$$logit(\pi_{ik}) = \delta_k + W_i'^{T}\gamma_k + \tilde{Q}_{ik}^{T}\lambda_k$$

Here δ_k are the intercepts and ${W_i}'^T\gamma_k$ are the effects of the predictor variables on the incomplete covariate being modelled. The last term of the linear predictor consists of

$$\widetilde{Q}_{ik}{}^T = \{Q_{i1}, \ldots, Q_{i4}\} \backslash \{Q_{ik}\}$$

with associated regression coefficients

$$\lambda_k = \{\lambda_{k1}, \ldots, \lambda_{k4}\} \backslash \{\lambda_{kk}\}.$$

and amounts to the effects of the three elements of $\boldsymbol{Q}$ not being modelled as the dependent variable. That is $\widetilde{Q}_{ik}{}^T\lambda_k = \sum_{(l \neq k)} \lambda_{kl} Q_{il}$.

Although the response model was fitted only to data from patients with a high distress score, we used data from all the patients to estimate the model for the incomplete covariates. This was for two reasons. Firstly, there was no reason only to use data from patients with a high distress score to predict the incomplete covariates, and it seemed sensible to increase confidence in the estimated parameters by basing these on the full sample. Secondly, the predictor analysis described in section 8.10 necessitated that covariate data be imputed for the full patient sample.

8.9 Results

Having specified vague normal prior distributions for the regression coefficients of the covariate model we used WinBUGS to fit model 8.1 for the responses jointly with model 8.2 for the incomplete covariates.

First, we fitted the model exactly as described in section 8.7 where associations with depression were assumed to be identical across the five cancer types. However, to allow for investigation of interaction effects with cancer type in the subsequent predictor analysis we also fitted a second model where the estimated associations with depression status were free to vary across cancer types. This was done by fitting the response model separately for each cancer type while the model for the

incomplete covariates was shared across cancer types and therefore unchanged from that set out in section 8.8. Cancer type remained an explanatory variable in the model for the incomplete covariates but was removed from the linear predictor in the response models.

Convergence was assessed using a combination of visual inspection of the iteration histories and formal convergence criteria evaluated by running simultaneous chains of the Gibbs sampler and monitoring the Gelman-Rubin statistic as described in Chapter 7. There were issues with convergence to a stationary distribution in the estimated intercept terms and regression coefficients for gender for the two response models relating to breast and gynaecological cancer. This is unsurprising since there were no males with gynaecological cancer and only very few males with breast cancer, none of whom were depressed (i.e. there were no events among males with breast cancer). Since the effect of gender was inestimable within those two cancer types the associated regression terms were removed from the two relevant models. Gender remained in the models for lung, genitourinary and lower gastrointestinal cancers. Based on assessment of model convergence we discarded the first 15,000 iterations from the Gibbs sampler (the burn-in) and based inference on the subsequent 10,000 iterations.

The results from fitting the model for the responses and missing covariate data are shown in Tables 8.3 and 8.4 respectively. The results in Table 8.3 are of somewhat limited clinical interest as the effects were estimated from a sample of patients with a clinic HADS score of 15 or more. However, it is reassuring that the results appear to be reasonably consistent across cancer types. The results from fitting the common model are also given for reference.

The results from fitting the model for the incomplete covariates are presented in Table 8.4. In contrast to the response model this model was fitted to the full sample of patients. The effect estimates were largely as one would expect. For example, patients were more likely to have been treated with curative intent if they had recently had surgery or radiotherapy, but not chemotherapy, if they were younger,

had breast cancer, and if they were diagnosed long ago. There was good symmetry in the estimated effects among the incomplete covariates themselves. None of the patients who were diagnosed more than three years ago had started cancer treatment in the past six months, and the associated odds were therefore infinitely smaller compared with the reference category. To avoid estimation problems we set these undefined odds equal to 0.0001 through the model specification.

Based on the 10,000 draws from the posterior predictive distribution for Y, the mean of the observed and predicted Y responses across all 21,151 patients was 0.096. This was the case both for the more general model and for the model where estimates were restricted to be identical across cancer types. This estimate of an overall depression prevalence in the sample of 9.6% was in good agreement with the previous estimate from section 8.5.

Table 8.3. Estimated parameters for the response model when fitted separately for each cancer type and when pooled across all cancer types *(Continued on next page)*.

	Multivariate model fitted separately within				
	Lung cancer (N=1,690)	Breast cancer (N=1,965)	Genito-urinary cancer (N=366)	Gynae-cological cancer (N=741)	Lower gastro-intestinal cancer (N=748)
Intercept	-1.04 (0.40)	-1.29 (0.35)	-2.06 (1.04)	-0.38 (0.49)	-0.40 (0.51)
Female gender	0.32 (0.14)	-	-0.77 (0.63)	-	-0.05 (0.21)
Age group					
< 50 years	0	0	0	0	0
50 to 59 years	-0.22 (0.33)	-0.24 (0.16)	-0.94 (0.64)	-0.02 (0.28)	-0.02 (0.37)
60 to 69 years	-0.57 (0.31)	-0.57 (0.17)	-0.63 (0.57)	-0.67 (0.27)	-0.68 (0.35)
≥ 70 years	-1.01 (0.32)	-1.39 (0.20)	-1.36 (0.57)	-1.23 (0.30)	-1.40 (0.37)
Health board*					
Lothian	0	0	0	0	0
Glasgow	0.20 (0.20)	0.14 (0.18)	0.58 (0.54)	-0.14 (0.32)	-0.35 (0.32)
Argyll & Clyde	-0.29 (0.29)	0.39 (0.18)	0.05 (0.54)	-0.63 (0.42)	-0.14 (0.41)
Lanarkshire	0.15 (0.23)	0.09 (0.20)	-0.48 (1.06)	-0.46 (0.32)	-0.80 (0.34)
Forth Valley	0.29 (0.28)	0.26 (0.31)	0.38 (0.55)	-0.01 (0.42)	-1.08 (0.43)
Tayside	-0.59 (0.33)	0.05 (0.27)	0.97 (1.12)	0.04 (0.43)	-0.60 (0.39)
Other†	0.01 (0.27)	0.17 (0.25)	0.49 (0.88)	-0.76 (0.32)	-0.86 (0.47)
Resident setting					
Urban	0	0	0	0	0
Small town	0.47 (0.26)	-0.30 (0.20)	0.08 (0.67)	-0.20 (0.37)	0.86 (0.38)
Rural	-0.07 (0.28)	-0.27 (0.21)	-0.49 (0.61)	0.06 (0.31)	0.56 (0.44)
Deprivation SIMD quintile score**					
1	0.35 (0.25)	0.39 (0.18)	2.29 (0.66)	0.22 (0.32)	0.65 (0.32)
2	0.47 (0.26)	0.31 (0.18)	1.65 (0.66)	0.12 (0.32)	0.23 (0.33)
3	0.16 (0.28)	0.49 (0.19)	1.71 (0.71)	0.05 (0.34)	0.34 (0.35)
4	-0.14 (0.32)	0.09 (0.19)	1.50 (0.67)	0.38 (0.37)	-0.28 (0.39)
5	0	0	0	0	0
Clinic HADS score (range 15-42)	0.13 (0.01)	0.18 (0.01)	0.14 (0.04)	0.19 (0.02)	0.16 (0.02)

Table 8.3. *(Continued on next page)*.

	Common multivariate model across cancer types (N=5,510)
Intercept	-0.92 (0.18)
Female gender	0.17 (0.11)
Age group	
< 50 years	0
50 to 59 years	-0.23 (0.11)
60 to 69 years	-0.63 (0.11)
≥ 70 years	-1.23 (0.12)
Health board*	
Lothian	0
Glasgow	0.11 (0.10)
Argyll & Clyde	0.04 (0.12)
Lanarkshire	-0.05 (0.12)
Forth Valley	0.06 (0.15)
Tayside	-0.11 (0.16)
Other†	-0.17 (0.14)
Resident setting	
Urban	0
Small town	0.09 (0.13)
Rural	-0.06 (0.13)
Deprivation SIMD quintile score**	
1	0.46 (0.11)
2	0.40 (0.12)
3	0.39 (0.13)
4	0.15 (0.13)
5	0
Clinic HADS score (range 15-42)	0.15 (0.01)

Table 8.3. *(Continued on next page)*.

	Multivariate model fitted separately within				
	Lung cancer (N=1,690)	Breast cancer (N=1,965)	Genito-urinary cancer (N=366)	Gynae-cological cancer (N=741)	Lower gastro-intestinal cancer (N=748)
Cancer type					
Lung	-	-	-	-	-
Breast	-	-	-	-	-
Genitourinary	-	-	-	-	-
Gynaecological	-	-	-	-	-
Lower gastrointestinal	-	-	-	-	-
Time since diagnosis					
< 3 months	0	0	0	0	0
3 to 12 months	0.32 (0.18)	0.69 (0.24)	0.03 (0.52)	0.70 (0.32)	0.84 (0.30)
1 to 3 years	0.60 (0.22)	1.30 (0.25)	-0.14 (0.54)	0.37 (0.36)	0.49 (0.33)
> 3 years	0.04 (0.30)	0.87 (0.24)	0.19 (0.52)	0.62 (0.37)	0.64 (0.38)
Started chemotherapy recently[£]	-0.01 (0.19)	-0.01 (0.25)	0.37 (0.77)	-0.24 (0.27)	-0.41 (0.38)
Started radiotherapy recently[£]	0.23 (0.19)	0.27 (0.24)	0.23 (0.74)	0.78 (0.37)	0.09 (0.53)
Received surgery recently[£]	0.48 (0.27)	0.15 (0.20)	0.14 (0.61)	-0.34 (0.29)	-0.12 (0.29)
Curative therapy objective	0.30 (0.18)	0.25 (0.26)	0.36 (0.37)	0.44 (0.26)	-0.07 (0.28)

Table 8.3. *(Continued)*.

	Common multivariate model across cancer types (N=5,510)
Cancer type	
Lung	0
Breast	-0.17 (0.12)
Genitourinary	-0.13 (0.17)
Gynaecological	0.06 (0.13)
Lower gastrointestinal	-0.26 (0.13)
Time since diagnosis	
< 3 months	0
3 to 12 months	0.52 (0.11)
1 to 3 years	0.63 (0.12)
> 3 years	0.49 (0.13)
Started chemotherapy recently[£]	-0.10 (0.11)
Started radiotherapy recently[£]	0.24 (0.12)
Received surgery recently[£]	-0.04 (0.11)
Curative therapy objective	0.31 (0.10)

Data are posterior means (standard deviations) of the log odds ratios. Model fitted with data from 1236, 1640, 280, 617, 591 and 4364 patients with observed depression status for each of lung, breast, genito-urinary, gynaecological and lower gastrointestinal cancers, and the common model respectively. *Health board (HB) where patient was resident when cancer was registered.[†] Ayrshire and Arran, Borders, Dumfries & Galloway, Fife, Grampian and Western Isles. [**] Scottish Index of Multiple Deprivation quintile score: 1=most deprived, 5=least deprived. [£] Treatment started in preceding six months.

Table 8.4. Estimated parameters from fitting the model for the incomplete covariates.

	Curative therapy objective (12% missing)	Started chemotherapy recently[£] (4% missing)	Started radiotherapy recently[£] (3% missing)	Received surgery recently[£] (6% missing)
Intercept	-0.90 (0.11)	-0.58 (0.10)	-2.36 (0.15)	-3.17 (0.11)
Curative therapy objective	-	-0.43 (0.07)	0.24 (0.07)	2.00 (0.07)
Started chemotherapy recently	-0.46 (0.06)	-	-0.89 (0.06)	-0.19 (0.06)
Started radiotherapy recently	0.47 (0.07)	-0.90 (0.06)	-	-0.69 (0.07)
Received surgery recently	2.00 (0.07)	-0.19 (0.06)	-0.62 (0.06)	-
Health board*				
Lothian	0	0	0	0
Glasgow	-0.74 (0.06)	0.14 (0.07)	-0.36 (0.08)	-0.09 (0.07)
Argyll & Clyde	-0.83 (0.07)	0.31 (0.08)	-0.46 (0.09)	-0.15 (0.09)
Lanarkshire	-0.67 (0.07)	0.17 (0.08)	-0.68 (0.10)	-0.15 (0.08)
Forth Valley	-0.43 (0.08)	0.45 (0.10)	-0.39 (0.11)	-0.30 (0.11)
Tayside	-0.23 (0.09)	0.05 (0.11)	-0.09 (0.11)	-0.33 (0.11)
Other[†]	-0.06 (0.08)	0.07 (0.09)	-0.14 (0.09)	-0.44 (0.09)
Age group				
< 50 years	0	0	0	0
50 to 59 years	-0.23 (0.09)	-0.44 (0.08)	0.02 (0.10)	0.39 (0.08)
60 to 69 years	-0.46 (0.08)	-0.80 (0.08)	0.00 (0.09)	0.58 (0.07)
≥ 70 years	-0.74 (0.08)	-1.22 (0.08)	-0.02 (0.10)	0.44 (0.08)
Deprivation SIMD quintile score**				
1	0.02 (0.06)	-0.15 (0.07)	0.06 (0.08)	-0.19 (0.07)
2	0.00 (0.06)	-0.10 (0.08)	0.18 (0.08)	-0.11 (0.07)
3	0.03 (0.07)	-0.14 (0.08)	0.13 (0.08)	-0.03 (0.08)
4	0.07 (0.07)	-0.12 (0.08)	0.04 (0.08)	-0.05 (0.08)
5	0	0	0	0
Cancer type				
Lung	0	0	0	0
Breast	2.68 (0.07)	-0.36 (0.08)	0.27 (0.08)	2.58 (0.08)
Genitourinary	0.70 (0.07)	-1.58 (0.14)	-1.25 (0.12)	0.69 (0.12)
Gynaecological	1.58 (0.07)	0.42 (0.08)	-1.03 (0.09)	1.73 (0.09)
Lower gastrointestinal	1.61 (0.06)	-0.17 (0.08)	-1.71 (0.11)	1.81 (0.08)
Time since diagnosis				
< 3 months	0	0	0	0
3 to 12 months	0.63 (0.06)	1.42 (0.06)	2.83 (0.08)	-0.21 (0.06)
1 to 3 years	1.52 (0.06)	-3.06 (0.17)	-0.65 (0.12)	-5.42 (0.16)
> 3 years	2.30 (0.08)	-812 (590.3)	-800 (602)	-793 (587)

Data are posterior means (standard deviations) of the log odds ratios. Parameter estimates are conditional on all other variables in the model. Model fitted with data from all 21,151 patients. [£] Treatment started in preceding six months. *Health board (HB) where patient was resident when cancer was registered.[†] Ayrshire and Arran, Borders, Dumfries & Galloway, Fife, Grampian and Western Isles. [**] Scottish Index of Multiple Deprivation quintile score: 1=most deprived, 5=least deprived. [£] Treatment started in preceding six months.

8.10 The predictor analysis

Having fitted the prediction model for the incomplete data as described, we were now in a position to address the clinical research questions. We wanted to estimate the depression prevalence within each of the five cancer types along with associations with patients' demographic and cancer characteristics. We did this by storing multiple independent imputed datasets from the Bayesian posterior predictive distribution for the missing data and analysing these using the SAS software.

The investigation of predictors was carried out using logistic regression to model associations with major depression. Could this not have been modelled jointly with the imputation models for the responses and incomplete covariates? The fact that the imputation model (model 8.1 - 8.2) was rather different from the substantive model (that used for the predictor analysis) made it technically challenging to fit these jointly since we would have had contradictory expressions for the expected value of Y_i within the same model specification. Besides, it seemed conceptually appealing to have disjoint models for the imputation and analysis stages of the problem.

As described in detail in Chapter 7 we generated 100 independent imputed datasets from the Bayesian model by storing the predicted values for the missing variables from every 100^{th} iteration of the Gibbs sampler. (In fact we stored 50 datasets from each of two simultaneously run chains and found no systematic differences in the scores generated by each chain.) The predicted data were combined with the observed data to form 100 complete datasets. Each of these datasets were analysed separately and the resultant estimates combined using multiple imputation rules.

From conversations with the clinical investigators we decided to include the following demographic and cancer characteristics in the logistic regression: gender, age group (<50 years; 50-59 years; 60-69 years; ≥70 years), time since diagnosis (<1 year; ≥1 year), recent cancer treatment (any of chemotherapy, radiotherapy or surgery started in the preceding six months [yes;no]), therapeutic objective (curative; palliative), resident setting (urban; small town; rural) and Scottish Index of Multiple Deprivation (SIMD) quintile score (range 1=most deprived to 5=least deprived).

Rather uniquely, due to the large sample size it was possible to analyse the data for differential subgroup effects across the five cancer types. We therefore extended the model to allow for interaction effects between cancer type and each of the remaining variables in the model (except for the effect of gender within gynaecological cancers). Ordinarily it would then be straightforward to assess the fit of the more general model over the reduced model through the likelihood ratio. However with multiple imputed data each dataset is different and it is not immediately obvious how evidence from each of the 100 likelihood ratio statistics should be combined. Alternatively the models can be contrasted using the Wald test to test the multidimensional hypothesis that the interaction terms are all equal to zero. This can be done by setting up $\mathbf{L\beta=0}$ where $\mathbf{L}$ is a matrix of linear contrasts designed to pick out the relevant elements of $\boldsymbol{\beta}$, the vector of regression coefficients. The test statistic is then tested against a chi-square distribution with degrees of freedom equal to the number of rows in $\mathbf{L}$. Technically the degrees of freedom for the reference distribution are also dependable on m, the number of imputed datasets used, although with $m=100$ we can rely on the asymptotic properties and resort to the usual chi-square reference distribution.

The evidence from the tests for differential subgroup effects are presented in table 8.5. Interestingly, the likelihood ratios of the general to the reduced model when derived for each of the imputed datasets separately were mostly significantly different from unity, but as is clear from table 8.5, when combining the evidence in the appropriate manner as described above there is no statistical evidence that such interaction effects exist. A single, common model across all five cancer types was consequently fitted. The model was estimated for each of the $m=100$ datasets and the results combined using Rubin's rules along with their multivariate generalisation for group tests of multidimensional parameter vectors. The results are shown in Table 8.6. (Despite the lack of statistical evidence for differential subgroup effects it might still be of clinical interest to have the analysis of associations available within each of the five common cancer types including the prevalence of major depression across

levels of the other covariates. The results from the analysis with the effects fitted separately for each cancer type are provided in Appendix A.)

The prevalence of depression was highest among patients with a lung cancer diagnosis (13.1%; 95% CI: 11.9 to 14.2%), followed by the almost exclusively female patient groups with gynaecological (10.9%; 95% CI: 9.8 to 12.1%) and breast cancer (9.3%; 95% CI: 8.7 to 10.0%). On the other hand, the almost exclusively male group of patients with genitourinary cancer had the lowest depression prevalence (5.6%; 95% CI: 4.5 to 6.7%) followed by the more mixed gender group of patients with lower gastrointestinal cancer (7.0%; 95% CI: 6.1 to 8.0%). These point and interval estimates were derived using multiple imputation rules.

Besides cancer type there were three other factors found to be predictive of major depression in the multivariate analysis. Female gender was predictive of major depression (OR: 1.49; 95% CI: 1.27 to 1.76). Younger age was a strong predictor with odds of depression 3.70 times higher (95% CI: 3.13 to 4.35) in the under 50s compared with those aged 70 years or more. Deprivation score was also a strong predictor with prevalence estimates ranging from 15% among those most deprived to 5% among the least deprived (OR: 2.91; 95% CI: 2.47 to 3.44).

Table 8.5. Tests for differential subgroup effects across cancer types.

Variable	Number of interaction terms / degrees of freedom	X^2 test statistic	p-value
Gender [#]	3	0.36	0.948
Age group	12	11.88	0.455
Time since diagnosis	4	6.44	0.169
Recent cancer treatment	4	5.04	0.283
Therapeutic objective	4	2.44	0.655
Resident setting	8	10.88	0.209
Deprivation SIMD quintile score	16	20.80	0.186
Global test of all interaction terms	51	56.10	0.290

Based on m=100 imputed datasets. Reference distribution used is chi-square. [#] Interaction with gender was not identified for gynaecological cancer.

Table 8.6. Prevalence and associations of major depression in outpatients with a cancer diagnosis.

	Total n (%)	Major depression Yes n (%)	Major depression No n (%)	Adjusted odds ratio (95% CI)	p-value
Total	21151 (100)	2031 (10)	19120 (90)		
Cancer type					**<.001**
Lower GI	3355 (16)	236 (7)	3119 (93)	1	
Breast	8461 (40)	788 (9)	7673 (91)	0.95 (0.79 to 1.14)	
Genitourinary	2009 (9)	113 (6)	1896 (94)	1.03 (0.79 to 1.35)	
Gynaecological	3010 (14)	329 (11)	2681 (89)	1.11 (0.90 to 1.36)	
Lung	4316 (20)	564 (13)	3752 (87)	1.81 (1.49 to 2.20)	
Gender					**<.001**
Male	6039 (29)	456 (8)	5583 (92)	1	
Female	15112 (71)	1575 (10)	13537 (90)	1.49 (1.27 to 1.76)	
Age group					**<.001**
<50 years	2521 (12)	400 (16)	2121 (84)	1	
50-59 years	4104 (19)	587 (14)	3517 (86)	0.88 (0.76 to 1.02)	
60-69 years	6820 (32)	626 (9)	6194 (91)	0.51 (0.44 to 0.59)	
≥70 years	7706 (36)	417 (5)	7289 (95)	0.27 (0.23 to 0.32)	
Time since diagnosis					**0.670**
<1 year	10694 (51)	1110 (10)	9584 (90)	1.03 (0.89 to 1.21)	
≥1 year	10457 (49)	921 (9)	9536 (91)	1	
Recent cancer treatment*					**0.257**
No	13642 (64)	1272 (9)	12370 (91)	1	
Yes	7509 (36)	759 (10)	6751 (90)	0.92 (0.79 to 1.07)	
Therapeutic objective					**0.586**
Palliative	5892 (28)	631 (11)	5261 (89)	1	
Curative	15259 (72)	1400 (9)	13860 (91)	0.96 (0.83 to 1.11)	
Resident setting					**0.103**
Urban	16689 (79)	1687 (10)	15002 (90)	1	
Small town	2001 (9)	169 (8)	1832 (92)	0.92 (0.76 to 1.10)	
Rural	2461 (12)	175 (7)	2286 (93)	0.83 (0.69 to 0.99)	
Deprivation SIMD quintile score**					**<.001**
1	4572 (22)	701 (15)	3871 (85)	2.91 (2.47 to 3.44)	
2	4259 (20)	482 (11)	3777 (89)	2.15 (1.80 to 2.55)	
3	3781 (18)	341 (9)	3440 (91)	1.73 (1.44 to 2.08)	
4	3731 (18)	254 (7)	3477 (93)	1.30 (1.07 to 1.59)	
5	4808 (23)	253 (5)	4555 (95)	1	

Missing data for the depression response, recent treatment and therapeutic objective were handled using multiple imputation with the reported frequencies averaged over the m=100 imputed datasets. Lower GI = lower gastrointestinal. *Any of chemotherapy, radiotherapy or surgery started in the preceding six months. **Scottish Index of Multiple Deprivation quintile score: 1=most deprived, 5=least deprived. Odds ratios are conditional on all other variables in the model.

8.11 Revisiting HADS<15 to indicate absence of depression

The above analysis took account of the many patients with a HADS score of 15 or more who failed to complete the subsequent interview for depression. However it was assumed that not a single one of the nearly 16,000 patients who scored less than 15 were depressed. The threshold of 15 was chosen because it was clinically convenient, not because of a natural divide reflecting fundamental qualities of the HADS. In fact, based on a previous study (Walker et al., 2007) we anticipated that approximately 1% of patients who scored low on the HADS were in fact depressed.

Figure 8.2 presents the distribution of HADS scores collected from the 21,151 patients who were included in the analysis sample, and for each HADS score, the number of patients who were actually diagnosed with major depression. Again, the numbers with actual diagnoses underestimate the true number with major depression because not all high scorers completed the depression interview.

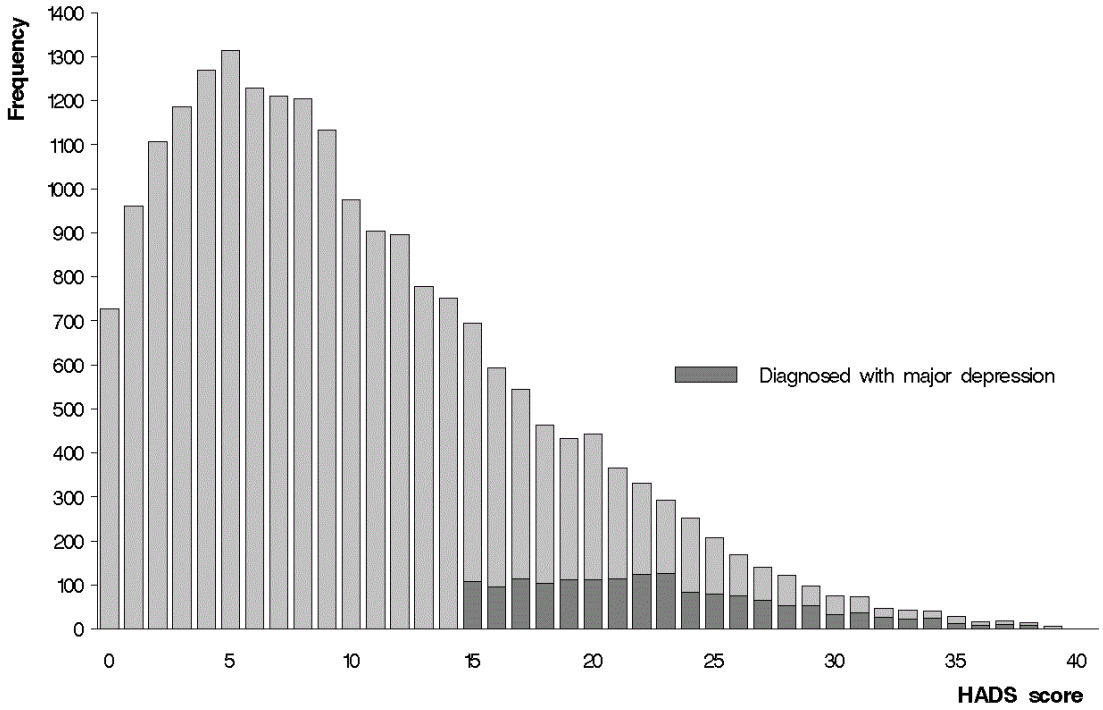


Figure 8.2. Distribution of HADS scores in the analysis sample (n=21,151) and number with a depression diagnosis (dark grey).

From looking at the distribution it is apparent that the majority of patients with a HADS score near the upper end of the scale are depressed and that the proportion of

depressed patients decreases with lower HADS scores. It also seems reasonable to suppose that this gradual trend would have continued into the lower range of HADS scores had the screening service also interviewed patients with HADS scores in this range.

When fitting the response model (model 8.1) to the high scorers we obtained estimates of the effect of the HADS score on the probability of depression. We found that the odds of depression were increased by a factor of 1.16 for each unit increase in the HADS. It is possible that this relationship between HADS score and depression applies to the lower end of the HADS scale as well. If so, we can use this to estimate the number of depressed patients who scored below 15 on the HADS. In fact we can improve on this even further by also taking into account the other covariate effects estimated in model 8.1.

We used the posterior distributions for the parameters estimated in model 8.1 to define the predictive distribution for the missing depression responses among patients with HADS scores less than 15. Overall, the model predicted that approximately 7.3% (95% CI: 5.9 to 8.7%) of those who scored low on the HADS were in fact depressed (Figure 8.3). If this were the case, the resultant overall prevalence estimate would be around 15.0% (95% CI: 13.9 to 16.2%).

The model prediction of the number of low scorers who were depressed was considerably higher than the 1% estimate from our previous study. Compared with other relevant studies included in our recent systematic review of the prevalence of depression in cancer patients, the overall estimate of 15% is also rather high. Together this suggests that the covariate effects that apply to the high HADS scorers cannot be applied to the whole population, and perhaps in particular that the relationship between HADS score and the log odds of depression cannot be extrapolated linearly in the way presumed.

Finally we note that there are alternative ways of modelling the depression scores at the lower end of the HADS which take into account the previous finding that only

1% of low scorers were depressed. One way to do so would be to fix the intercept of the prediction model to ensure a common probability of 1%. Alternatively we could use the prior distribution on the intercept to restrict departure from a common probability of 1% in a way that accommodates our confidence in this previous finding. However we will not pursue this any further here.

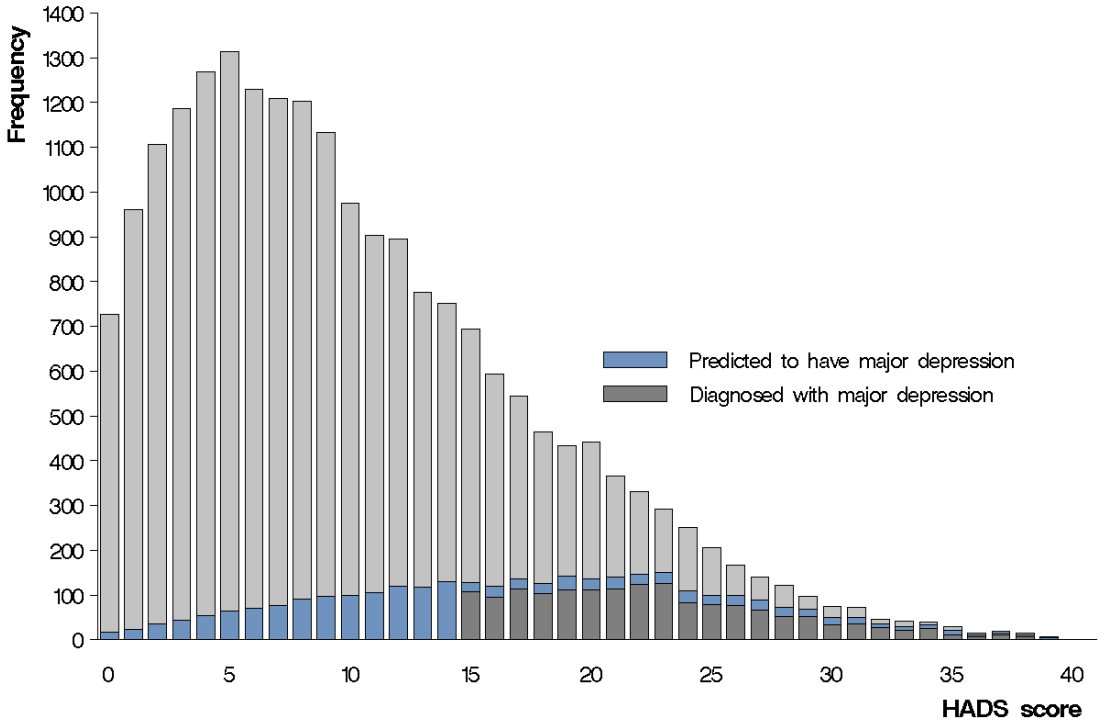


Figure 8.3. Distribution of HADS scores in the analysis sample (n=21,151) and number with observed and predicted depression.

8.12 Further sensitivity analysis

The results presented in section 8.10 were derived under a MAR assumption. But perhaps it was unreasonable to assume that the levels of depression among patients who failed to complete the depression interview were similar to those observed among people who completed the interview.

Modelling the refusal mechanism

Approximately one third of patients who scored high on the HADS and who were not interviewed for depression had refused the offer of an interview. It seemed reasonable to suppose that suffering from a major depressive episode would increase the probability of a patient refusing. We expected that the odds of refusal might be

twice as high in patients with depression, and that it would be no more than four times as high.

To assess the sensitivity of the findings to these altered assumptions we modelled the missingness mechanism jointly with the model for the responses and missing covariates (model 8.1-8.2) in patients with HADS≥15. We took a similar approach as that outlined in section 7.10 by specifying a model for the probability of patient *i* refusing interview conditional on that patient having been offered an interview for depression:

$$R_i \sim Bernoulli(\pi_{Ri})$$

$$logit(\pi_{Ri}) = \alpha_R + \beta_R Y_i$$

Here $R_i = 1$ if the patient refused and $R_i = 0$ if the patient accepted the offer of interview. π_{Ri} is the conditional probability of refusal given the patient did not fail to interview for other reasons. α_R, the log odds of refusal in patients who were not depressed, was estimated from the data once β_R, the increase in log odds of refusal among patients who were depressed, was fixed. The model was fitted jointly with model 8.1-8.2 first setting $\beta_R = log(2)$ and secondly with $\beta_R = log(4)$.

The first scenario under which the odds of refusal were assumed twice as high in patients with depression yielded a value for $\alpha_R = -2.61$. Under this model, conditional on the patients having been offered an interview, the probability of refusal in patients who were depressed was therefore $e^{(-2.61+log(2))} = 0.148$ compared with $e^{(-2.61)} = 0.074$ in those who were not depressed. Under the second scenario where the odds of refusal were assumed four times as high in patients with depression, we found that $\alpha_R = -3.03$, and the corresponding probabilities of refusal were *0.193* and *0.048* in patients who were depressed and not depressed respectively.

As expected the two models resulted in depression prevalence estimates that were somewhat increased (Table 8.7). We also repeated the predictor analysis described in

section 8.10 using multiple imputed datasets from these two selection models; the estimated associations were very similar indeed to the estimated associations under the main model.

Table 8.7. Depression prevalence estimates (95% CIs) under the two alternative MNAR assumptions for the refusal mechanism.

Cancer type	β_R=log(2)	β_R=log(4)
Lung	14.1% (13.0% to 15.3%)	15.0% (13.8% to 16.2%)
Breast	9.5% (8.8% to 10.1%)	9.7% (9.0% to 10.4%)
Genitourinary	5.9% (4.7% to 7.0%)	6.1% (5.0% to 7.2%)
Gynaecological	11.1% (9.9% to 12.3%)	11.3% (10.1% to 12.5%)
Lower gastro-intestinal	7.3% (6.4% to 8.3%)	7.6% (6.6% to 8.6%)

MNAR offset

In a second sensitivity analysis we considered two scenarios again concerning patients who had scored high on the HADS without completing the depression interview. We repeated the main analysis under the assumption that (a) these patients had half the odds of being depressed and (b) that they had twice the odds of being depressed compared with patients who had completed the interview.

We incorporated these assumptions by adding the term δM_i to model 8.1, where $M_i=1$ (*0*) when Y_i was missing (observed). Scenarios (a) and (b) corresponded to offset values of $\delta=log(0.5)$ and $\delta=log(2)$ respectively.

Approximately one fifth of patients who had scored high on the HADS did not complete the depression interview. Under scenario (a), the prevalence in this group was estimated to be 26.5%, somewhat less than the 36.6% observed among those who completed the interview. Under scenario (b) however, the equivalent estimated prevalence among the unobserved group was 51.0%.

The prevalence estimates under these two scenarios are shown in Table 8.8 for each of the five cancer types. As expected, the prevalence estimates were somewhat lower under scenario (a) and higher under scenario (b) compared with the main analysis, however the relative differences between the cancer types did not change much.

Again, the predictor analysis was repeated; there were no substantial differences in the estimated associations with major depression.

Table 8.8. Depression prevalence estimates (95% CIs) under two alternative MNAR assumptions for the δ-offset.

Cancer type	*δ=log(0.5)*	*δ=log(2)*
Lung	11.9% (10.8% to 13.0%)	14.6% (13.4% to 15.8%)
Breast	8.8% (8.2% to 9.5%)	9.8% (9.1% to 10.5%)
Genitourinary	5.2% (4.2% to 6.3%)	6.2% (5.0% to 7.3%)
Gynaecological	10.5% (9.3% to 11.6%)	11.5% (10.3% to 12.7%)
Lower gastro-intestinal	6.6% (5.7% to 7.5%)	7.6% (6.6% to 8.6%)

8.13 Discussion

The study aimed to determine the prevalence of depression in patients with a cancer diagnosis attending outpatient oncology clinics, and to identify demographic and clinical characteristics that were predictive of depression.

We found that the depression prevalence differed markedly between cancer types, although some of the differential effects were attributable, at least in part, to gender differences. Depression tended to be more prevalent in female dominated cancers (breast: 9%; gynaecological: 11%) and less so in more male dominated cancers (genitourinary: 6%; lower gastrointestinal: 7%). But it was the mixed gender lung cancer group who were most likely to suffer from major depression (13%). Common to all cancer types, the study found that female gender, younger age and greater deprivation were highly associated with an increased risk of depression. The time since diagnosis and treatment, and the therapeutic objective, were not predictive of depression.

In addition to the fully adjusted odds ratios we reported prevalence estimates for subgroups of patients defined by their gender, age, time since diagnosis and treatment, therapeutic objective, resident setting and level of deprivation. Because of the unusually large dataset it was also possible to present this information separately within each of the five common cancer types (presented in Appendix A). To our knowledge, the prevalence and associations of depression in cancer patients have not previously been studied in this depth and breadth.

8.13.1 Limitations

Because of the two-stage screening procedure and missingness both in the response and covariate data this was not a straightforward prevalence estimation problem, and the analysis consequently rested on a number of assumptions.

However, we took a principled approach to the analysis. The main analysis was conducted under a MAR assumption: we estimated the posterior predictive distribution for the missing data in WinBUGS and used this to generate multiple independent imputed datasets for subsequent analysis in SAS. Secondly in a number of sensitivity analyses we used the same approach to generate multiple imputed data under various MNAR scenarios re-estimating the model under two different types of informative missingness, and choosing values to quantify the departure from MAR that spanned the set of plausible values.

Compared with the main analysis, the results from the sensitivity analysis yielded prevalence estimates that were a few percentage points higher or lower depending on the direction and size of the departure from MAR. However, the finding that depression was most prevalent in lung cancer patients followed by gynaecological, breast, lower gastrointestinal and genitourinary cancer patients, in that order, remained unchanged in the sensitivity analysis. Female gender, younger age and greater deprivation were strongly associated with increased risk of depression under all scenarios. The consistency in these results provided added confidence in the validity of the study findings.

8.13.2 Other directions

With more time we would have liked to have carried out further analyses: some to investigate in more detail the robustness of the findings from the main analysis, and others to widen the external validity of the study.

For example, to improve prediction of the missing depression responses it would have been possible to have extended the response model (model 8.1) to also include

patients' HADS scores from subsequent clinic visits. Alternatively we could have modelled subsequent depression responses jointly with the depression response of interest to improve the predictions. This would be similar to the use of auxiliary data points in Chapter 7. In addition to symptoms of psychological distress, the screening service also asked about symptoms of pain, fatigue, disturbed sleep and nausea/ vomiting. It is possible that these could have been used to improve predictions of the missing depression responses too, although the explanatory power of these would presumably be limited after adjusting for a patient's HADS score.

It would also have been interesting to have repeated the analysis using data collected from patients' first clinic appointment only, ignoring any subsequent appointments. This is the second of the three options that are listed in section 8.3 for selecting the analysis sample. Clearly this selection strategy would have resulted in a substantial degree of missing HADS and depression responses which would have had to have been handled in the analysis.

To broaden the external validity of the analysis we could have widened the analysis sample to include all 31,329 patients with appointments, including those who never completed the HADS. However, predictions from these patients would have been particularly unreliable because the linked research database prepared by NHS Scotland Information Services Division contained no information on these patients.

In a final extension, one could consider all patients living in Scotland with a cancer diagnosis, including those who did not have any recorded clinic appointments. Such an extrapolation would rely mostly on untestable assumptions but is worth a thought nonetheless since it is likely that the findings from this study will be interpreted in this widest of contexts.

9 CONCLUSION

9.1 Summary of findings

The observational symptoms data routinely collected by the oncology outpatient depression screening service in Scotland are possibly the most extensive collection of such data in the world. As is generally the case with routinely observed healthcare data, the limitations to these data were many. Nonetheless, we wondered whether with the right analysis approach these data might offer a unique opportunity for observational study on an unprecedented scale into this population.

Meanwhile the POD Study was conducted to investigate the progression of distress symptoms over time in patients who had presented at a screened clinic with signs of significant distress but without meeting criteria for major depression. The study was undertaken in part to inform the design of a potential trial in this population. It was unknown to what extent such distress symptoms were likely to persist over time and to require treatment. The study showed that the distress prevalence dropped quickly following enrolment into the study, but remained rather constant thereafter and that around 38% of patients were still suffering from significant distress seven months after their clinic appointment. Distress status at one month follow-up was a strong predictor of persistent distress at seven months. There was also evidence that patients recently treated with radio or chemotherapy were less likely to remain distressed at seven months. A potential trial should aim to enrol patients who exhibit symptoms of significant distress at clinic appointment and again one month later.

Having identified a subsample of patients in the screening database that satisfied the POD Study inclusion criteria we used the observational symptom data from this sample to construct a dataset which matched the structure of the POD Study. Most patients represented in the screening data attended appointments at a frequency that was much less than the frequency with which follow-up data were accrued in the POD Study. As a consequence, there was a substantial amount of missing data. We developed a model to predict the incomplete longitudinal responses (along with the limited incomplete covariates) under both MAR and MNAR mechanisms using Bayesian methods and used this to produce multiple imputations for further analysis

in SAS. We found as with the POD Study that distress levels dropped initially and remained moderately constant with around 46% still distressed at seven months. Likewise, distress status at one month follow-up was seen to be a strong predictor of persistent distress also in this analysis. But there were also systematic departures from the POD Study findings. The mean distress and prevalence estimates were consistently higher in the screening data, and there were some differences in the results from the predictor analysis as well. Lastly, we used the observational symptom data from the screening service to conduct a comprehensive analysis to estimate the prevalence and associations of depression in this population. We estimated the depression prevalence in the five most common types of cancer and found that this varied from 13.1% in lung cancer patients to 5.6% in patients with genitourinary cancer. Female gender, deprivation and younger age were found to be strongly associated with depression. This will likely be the largest study in world to answer these questions.

A recent systematic review of the prevalence of depression in cancer patients found that estimates vary widely due to small and poorly executed studies and depending on the clinical setting and diagnostic criteria used to identify patients who are depressed (Walker et al., 2012). The review, which included only studies that met basic quality criteria and used diagnostic interviews to determine caseness, found prevalence estimates in mixed cancer outpatients ranging from 5% to 16% and reported generally higher prevalence estimates from studies in palliative care settings. A separate review of major depression in breast cancer patients (Fann et al., 2008) found rates of depression from about 10% to 25%. This report found higher rates in studies that based findings on screening instruments rather than diagnostic interviews.

Two US studies included in a review by Mitchell et al. (2001) and which were rated as low quality reported depression prevalence rates among patients with gynaecological cancers around 23% (Evans et al., 1986; Golden et al., 1991). Two other small studies from the review also rated as low quality found very different rates of depression in lung cancer patients (2%) and head and neck or lung cancer

patients (20%) in Canada and Australia respectively (Ginsburg et al., 1995; Kangas et al., 2005)

The review also reported a pooled depression prevalence in mixed cancer in- and outpatients from non-palliative-care settings of 16.3% (95% CI: 13.4 to 19.5). This was markedly higher than the prevalence estimates derived in Chapter 8. However the review reported large heterogeneity in the estimates and found strong evidence of publication bias with few small studies reporting a low prevalence. Finally this review found many studies reporting the prevalence of depression individually for breast cancer patients and for mixed cancer groups, but found few that reported individually on other cancer groups as is done in Chapter 8.

9.2 Context

Where purpose-designed, prospective studies are not feasible, observational healthcare data can be used to study safety and effectiveness outcomes from medical interventions and for epidemiological research to study associations between exposures and outcomes. Furthermore, with the right analysis approach and a good background understanding we believe that pilot work with observational healthcare data could be used to inform future trial designs. In Chapter 7 the POD Study eligibility criteria were applied to the observational data to identify an analysis sample from the screening data that mirrored the POD Study sample at the time of enrolment. In a similar manner, Danaei et al. (2013) use observational data to emulate a randomised controlled trial to estimate the effect on coronary heart disease of initiating treatment with statins. The authors achieve surprisingly sound results and conclude that meaningful analysis of observational healthcare data requires background knowledge, high quality information and appropriate analytical methods. Equally, Overhage & Overhage (2013) write that analysis with observational data can produce valuable insight if careful attention is given to the limitations and the particular characteristics of the data. Ryan (2013) concludes that inter-disciplinary collaboration is key to advancing clinical research with observational healthcare data and that statisticians should play an integral role in ensuring the sound use and interpretation of such data.

9.3 Limitations and other directions

The present project is entitled *Analysis of routinely collected repeated patient outcomes*. Although we chose to study this in the context of missing data there are clearly many other ways that the topic can be approached. In the central analysis of Chapter 7 we forced an artificial temporal structure onto the irregularly observed data and regarded the discrete time points as either observed or missing. The chapter offers just this one strategy for handling the irregular spacing of the measurements. Other analysis approaches might have resulted in a more efficient use of the data. As an alternative to this setup we could have analysed the data in continuous time by fitting parametric curves to the available data, thus avoiding discarding data where multiple measurements existed within a single time window. Instrumental variables methods are often used for causal effects estimation in econometrics in attempts to control for unmeasured confounding. We did not find these methods directly relevant to the present research since the concern is not one of estimation of central causal effects. Moreover, these methods rely on the presence of instruments that are associated with the explanatory variable of interest but not with the outcome modelled. These are strong assumptions that cannot be empirically verified.

The work described in this thesis is the result of a dynamic process. The project has evolved along with my understanding of the topic. Inevitably some chapters were written earlier on in the process; given the chance again there are a number of things that I would do differently. The review of missing data methodology in Chapter 3 puts disproportionate emphasis on multiple imputation. However much of the content placed under this heading is generic to analysis with missing data. I would also have liked to have included a detailed account of the EM algorithm and its applications, a section on the special cases when closed-form derivation of incomplete data-likelihoods is possible, and a more thorough treatment of methods for analysis with non-ignorable missing data. I would also have provided a review of MCMC methods and given an account of Gibbs sampling. In Chapter 7 and 8 the likelihood for the four incomplete covariates was specified in terms of a chain of fully conditional models, each conditional on the other three jointly modelled incomplete covariates.

A less complicated alternative to this specification as suggested by Ibrahim et al. (2005) would be to model the joint distribution of the incomplete covariates Q using the following hierarchical factorisation:

$$f(Q1, Q2, Q3, Q4) = f(Q4|Q3, Q2, Q1)\, f(Q3|Q2, Q1)\, f(Q2|Q1)\, f(Q1)$$

For comparison, the main model in Chapter 7 was refitted using this alternative model specification; the results of this preliminary analysis suggested that computation time was reduced by about one sixth. The resulting posterior mean parameter estimates and standard deviations were practically identical to those from the original model with the exception of the covariate model parameters belonging to the three logistic regressions where the linear predictors had been reduced.

The work presented in chapter 7 would have benefitted from reanalysis under alternative model formulations. We modelled the responses in a hierarchical model using univariate normal distributions at each time point that were linked via a random intercept. We could instead have specified the joint multivariate normal distribution explicitly. Alternatively the flexibility offered by WinBUGS could have been exploited to condition the outcome at any one time point on a function of the adjoining outcomes, for example in a way that would take into account their relative proximity in time. Future research could also explore the role of prior distributions in order to take full advantage of the Bayesian approach. The model presented could have been supplemented with assessments of model diagnostics at various stages of the development. Furthermore, the analysis with the screening data in Chapter 7 could have been extended to also include estimation of the distress trajectories and the comparative utility of confirmatory distress measurements at one, two and four months follow-up as was done with the POD Study sample (sections 5.7.5 to 5.7.7).

For the prediction of depression outcomes in Chapter 8 it would have been fitting to have modelled outcomes from other time points jointly with the outcome of interest in the prediction model to make full use of the longitudinal nature of the data in a way similar to the use of auxiliary time points in Chapter 7. Finally the screening data included complete details on the antidepressants and psychological treatments

received by patients identified with major depression; the inclusion of these data will enrich the research presented in Chapter 8.

9.4 Implications

The question that motivated this project was whether, and if so how, meaningful analysis using observational healthcare data can be achieved. Clearly there is no answer in general to this question since no analysis will suit all problems, and no amount of statistical ingenuity can compensate for poor data. However, of course there are strategies and methodological approaches that are better suited to such analysis than others depending on the context and data.

We studied the question within the confines of the observational symptom screening data and the POD Study research aims and asked whether, through analysis with the screening data, we could reproduce the clinical study findings. It now seems that this was rather a tough test because the temporal structure of the screening data was not ideally suited to address the aims of the POD Study. As a consequence the analysis rested to a large degree on model predictions. In spite of this the overall findings from the two analyses were much the same, and there were no directly contradicting results. Still, there were systematic differences in the results as described above and one would probably not ordinarily have recommended using these observational data for the research aims pursued in the POD Study. On the other hand the mismatch between the screening data and the POD Study research aims served to illustrate some of the central challenges that arise when using observational data for research.

So, were we able to reproduce the POD Study findings from analysis with the observational screening data? Mostly, but not excellently. In contrast, the cross-sectional design of the study presented in Chapter 8 was probably better suited to analysis with the screening data. Setting out the motivation for the present project in section 1.2 we wondered about the comparative advantages of a vast but unstructured set of observational data versus a much smaller but deliberately designed clinical study. At least here the answer was clear: Thanks to the routinely collected symptom data we were able to study depression and its associations on an unprecedented scale

which we could not have achieved in a prospective clinical study. In conclusion, while it is not possible to provide a general answer to the question *can observational healthcare data be used for research*, the work presented in this thesis demonstrates that analysis with such data can potentially be advanced considerably with the use of flexible and innovative modelling techniques now made practicable with modern computing power.

9.5 References

Barnett AG, van der Pols JC, Dobson AJ. 2005, "Regression to the mean: what it is and how to deal with it", *International Journal of Epidemiology*, vol. 34, no. 1, pp. 215-220.

Beath KJ, Dobson AJ. 1991, "Regression to the mean for nonnormal populations", Biometrika, vol. 78, no. 2, pp. 431-435.

Blisker D, Goldner EM. 2002, "Routine outcome measurement by mental health-care providers: is it worth doing?", *Lancet*, vol. 360, pp. 1689-90.

Brooks SP, Gelman A. 1998, "Alternative methods for monitoring convergence of iterative simulations", *Journal of Computational and Graphical Statistics*, vol. 17, pp. 434-455.

Carpenter J, Kenward M, Evans S, White I. 2004, "Last observation carry-forward and last observation analysis", *Statistics in Medicine*, vol. 23, pp. 3241–3242.

Carrigan G, Barnett AG, Dobson AJ, Mishra G (2007). Compensating for missing data from longitudinal studies using WinBUGS. *Journal of Statistical Software*, Vol. 19 7, pp 1-17.

Curcin V, Bottle A, Molokhia M, Millett C, Majeed A. 2010, "Towards a scientific workflow methodology for primary care database studies", *Statistical Methods in Medical Research*, vol. 19, pp. 378-93.

Danaei G, Rodríguez LAG, Cantero OF, Logan R, Hernán MA. 2013, "Observational data for comparative effectiveness research: An emulation of randomised trials of statins and primary prevention of coronary heart disease", *Statistical Methods in Medical Research*, vol. 22, pp. 70-96.

Das P, Mulder PHG. 1983, “Regression to the Mode”, *Statistica Neerlandica*, vol. 37, no. 1, pp. 15-20.

Davies HTO, Crombie I K. 1997, “Interpreting health outcomes”, *Journal of Evaluation in Clinical Practice*, vol. 3, pp. 187-199.

Davis CE. 1976, “The effect of regression to the mean in epidemiologic and clinical studies”, *American Journal of Epidemiology*, vol. 104, no. 5, pp. 493-498.

Dempster AP, Laird NM, Rubin DB. 1977, “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38.

Diggle P, Kenward MG. 1994, “Informative Drop-out in Longitudinal Data Analysis”, *Journal of the Royal Statistical Society*, vol. 43, pp. 49-93.

Diggle PJ, Heagerty P, Liang KY, Zeger SL. 2002, "Analysis of Longitudinal Data", 2nd edition, Oxford: Oxford University Press.

Epstein AM. 1990, “The outcomes movement: will it get us where we want to go?”, *New England Journal of Medicine*, vol. 323, pp. 266-70.

Evans DL, McCartney CF, Nemeroff CB, Raft D, Quade D, Golden RN, Haggerty JJ, Holmes V, Simon JS, Droba M. 1986, “Depression in women treated for gynecological cancer: clinical and neuroendrocrine assessment”, *American Journal of Psychiatry*, vol. 143, pp. 447–452.

Fann JR, Thomas-Rich AM, Katon WJ, Cowley D, Pepping M, McGregor BA, Gralow J. 2008, “Major depression after breast cancer: a review of epidemiology and treatment”, *General Hospital Psychiatry*, vol. 30, pp. 112-126.

First MB, Spitzer RL, Gibbon M, Williams JBW. 1999, "Structured Clinical Interview for DSM-IV Axis I Disorders", Biometrics Research Department, New York State Psychiatric Institute: New York.

Fitzmaurice GM, Laird NM, Ware JH. 2004, "Applied Longitudinal Analysis", New Jersey: John Wiley.

Fitzmaurice GM. 2003, "Methods for Handling Dropouts in Longitudinal Clinical Trials", *Statistica Neerlandica*, vol. 57, pp. 75-99.

Galton F. 1886, "Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute*, vol. 15, pp. 246-263.

Gelman A. 2006, "Prior distributions for variance parameters in hierarchical models", *Bayesian Analysis*, vol. 1, pp. 515-534.

Gilbert R, Fluke J, O'Donnell M, Gonzalez-Izquierdo A, Brownell M, Gulliver P, Janson S, Sidebotham P. 2012, "Child maltreatment: variation in trends and policies in six developed countries", *Lancet*, vol. 379, pp. 758-72.

Ginsburg ML, Quirt C, Ginsburg AD, MacKillop WJ. 1995, "Psychiatric illness and psychosocial concerns of patients with newly diagnosed lung cancer", *Canadian Medical Association Journal*, vol. 152, pp. 701–708.

Golden RN, McCartney CF, Haggerty JJ, Raft D, Nemeroff CB, Ekstrom D, Holmes V, Simon JS, Droba M, Quade D, Fowler WC, Evans DL. 1991, "The detection of depression by patient self-report in women with gynecologic cancer", *The International Journal of Psychiatry in Medicine*, vol. 21, pp. 17–27.

Greenhalgh, J. 2009, "The applications of PROs in clinical practice: what are they, do they work, and why?", *Qual Life Res*, vol. 18, pp. 115-123.

He Y, Zaslavsky AM, Landrum MB, Harrington DP, Catalano P. 2010, "Multiple imputation in a large-scale complex survey: a practical guide", *Statistical Methods in Medical Research*, vol. 19, pp. 653-70.

Hogan JW, Lancaster T. 2004, "Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies", *Statistical Methods in Medical Research*, vol. 13, pp. 17-48.

Holm Hansen C, Walker J, Thekkumpurath P, Kleiboer A, Beale C, Sawhney A, Murray G, Sharpe M. 2013, "Screening medical patients for distress and depression: does measurement in the clinic prior to the consultation overestimate distress measured at home?", *Psychological Medicine*, Published Online First: January 2013, doi:10.1017/S0033291712002930.

Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. 2005, "Missing-data methods for generalized linear models: A comparative review", Journal of the American Statistical Association, vol. 100, pp. 332-346.

Kangas M, Henry JL, Bryant RA. 2005, "The course of psychological disorders in the 1st year after cancer diagnosis", *Journal of Consulting and Clinical Psychology*, vol. 73, pp. 763–768.

Kenward MG, Carpenter J. 2007, "Multiple imputation: current perspectives", Statistical Methods in Medical Research, vol. 16, pp. 199-218.

Lagarde M. 2012, "How to do (or not to do) . . . Assessing the impact of a policy change with routine longitudinal data", *Health Policy and Planning*, vol. 27, pp. 76–83.

Laird NM. 1988, "Missing data in longitudinal studies", *Statistics in Medicine*, vol. 7, pp. 305-315.

Le HV, Beach KJ, Powell G, Pattishall E, Ryan P, Mera RM. 2013, "Performance of a semi-automated approach for risk estimation using a common data model for longitudinal healthcare databases", *Statistical Methods in Medical Research*, vol. 22, pp. 97-110.

Leaf RC, DiGiuseppe R, Mass R, Alington DE. 1993, "Statistical methods for analysis of incomplete clinical service records: concurrent use of longitudinal and cross-sectional data", *Journal of Consulting and Clinical Psychology*, vol. 61(3), pp. 495-505.

Lee KJ, Carlin JB. 2010, "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation", *American Journal of Epidemiology*, vol. 171(5), pp. 624-32.

Li L, Shen C, Li X, Robins JM. 2013, "On weighting approaches for missing data", *Statistical Methods in Medical Research*, vol. 22, pp. 14-30.

Li X, Shen C. 2013, "Linkage of patient records from disparate sources",*Statistical Methods in Medical Research* , vol. 22, pp. 31-38.

Liang KY, Zeger SL. 1986, "Longitudinal data analysis using generalized linear models", *Biometrika*, vol. 73, pp. 13-22.

Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. 2004, "Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma", *Lancet*, vol. 363, pp. 1147-54.

Little RJA, Rubin DB. 2002, "Statistical analysis with missing data", 2nd edition, New York: John Wiley.

Little RJA. 1988, "Missing-data adjustments in large surveys", *Journal of Business & Economic Statistics*, vol. 6, pp. 287-296.

Little RJA. 1995, “Modelling the Drop-Out Mechanism in Repeated-Measures Studies”, *Journal of the American Statistical Association*, vol. 90, pp. 1112-1121.

Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. 2013, “The BUGS book. A practical introduction to Bayesian analysis”, Chapman & Hall/CRC Texts in Statistical Science Series.

Macdonald AJD. 2002, “The usefulness of aggregate routine clinical outcomes data: The example of HoNOS65+”, *Journal of Mental Health*, vol. 11, pp. 645-56.

Mallinckrodt CH, Kaiser CJ, Watkin JG, Molenberghs G, Carroll RJ. 2004, “The effect of correlation structure on treatment contrasts estimated from incomplete clinical trial data with likelihood-based repeated measures compared with last observation carried forward ANOVA”, *Clinical Trials*, vol. 1, pp. 477-489.

McDonald SA, Hutchinson SJ, Bird SM, Mills PR, Dillon J, Bloor M, Robertson C, Donaghy M, Hayes P, Graham L, Goldberg DJ. 2009, “A population-based record linkage study of mortality in hepatitis C-diagnosed persons with or without HIV coinfection in Scotland”, *Statistical Methods in Medical Research*, vol. 18, pp. 271-83.

Meng XL. 1994, “Multiple-imputation inferences with uncongenial sources of input”, *Statistical Science*, vol. 9, pp. 538–558.

Mitchell AJ, Chan M, Bhatti H, Halton M, Grassi L, Johansen C, Meader N. 2011, “Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: a meta-analysis of 94 interview-based studies”, *Lancet Oncology*, vol. 12, pp. 160-174.

Molenberghs G, Kenward MG. 2007, “Missing data in clinical studies”, Chichester: John Wiley.

Overhage JM, Overhage LM. 2013, “Sensible use of observational clinical data”, *Statistical Methods in Medical Research*, vol. 22, no. 1, pp. 7-13.

Pinto-Meza A, Serrano-Blanco A, Penarrubia MT, Blanco E, Haro JM. 2005, “Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone?”, *Journal of general internal medicine*, vol. 20, no. 8, pp. 738-742.

Robins JM, Gill RD. 1997, “Non-response models for the analysis of non-monotone ignorable missing data”, *Statistics in Medicine*, vol. 16, pp. 39-56.

Rose MR, Bezjak A. 2009, “Logistics of collecting patient-reported outcomes (PROs) in clinical practice: an overview and practical examples”, *Qual Life Res*, vol. 18, pp. 125-136.

Rosenbaum PR, Rubin DB. 1983, “The central role of the propensity score in observational studies for causal effects”, *Biometrika*, vol. 70, pp. 41-55.

Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. 2012, “Data from clinical notes: a perspective on the tension between structure and flexible documentation”, *Journal of the American Medical Informatics Association*, vol. 18, pp. 181-6.

Rubin DB. 1976, “Inference and missing data”, Biometrika, vol. 63, pp. 581-92.

Rubin DB. 1977, “Formalizing subjective notions about the effect of nonrespondents in sample surveys”, *Journal of the American Statistical Association*, vol. 72, pp. 538-543.

Rubin DB. 1978, “Multiple imputations in sample surveys”, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20-28.

Rubin DB. 1987, “Multiple imputation for nonresponse in surveys”, New York: John Wiley.

Ryan P. 2013, “Statistical challenges in systematic evidence generation through analysis of observational healthcare data networks”, *Statistical Methods in Medical Research*, vol. 22, pp. 3-6.

SAS OnlineDoc® 9.1.3. Cary, NC: SAS Institute Inc.

SAS software, Version 9.1. Copyright © 2002-2003 SAS Institute Inc., Cary, NC, USA.

Schafer JL, Graham JW. 2002, “Missing Data: Our View of the State of the Art”, *Psychological Methods*, vol. 7, No. 2, pp. 147-177.

Schafer JL. 1997, “Analysis of incomplete multivariate data”, London: Chapman & Hall.

Schafer JL. 1999, “Multiple imputation: a primer”, *Statistical Methods in Medical Research*, vol. 8, pp. 3-15.

Senn S. 2009, “Three things that every medical writer should know about statistics”, *The Journal of the European Medical Writers Association*, vol. 18, no. 3, pp. 159-162.

Sharma N, Holm Hansen C, O'Connor M, Walker J, Kleiboer A, Murray G, Espie C, Story D, Sharpe, M. 2011, “Sleep problems in cancer patients: prevalence and association with distress and pain”, Psycho-oncology. doi: 10.1002/pon.2004.

Sithole JS, Jones PW. 2003, “Analysis of a Bayesian repeated measures model for detecting differences in GP prescribing habits”, *Statistical Methods in Medical Research*, vol. 12, pp. 475-87.

Sparapani R. 2004, "Some SAS macros for BUGS/WinBUGS data", *The ISBA bulletin*, vol. 11, pp. 8-10.

Spiegelhalter D, Thomas A, Best N, Lunn D (2003). WinBUGS User Manual, Version 1.4. MRC Biostatistics Unit, Cambridge.

Spratt M, Carpenter J, Sterne AC, Carlin JB, Heron J, Henderson J, Tilling K. 2010, "Strategies for Multiple Imputation in Longitudinal Studies", *American Journal of Epidemiology*, vol. 172, No. 4, pp. 478-487.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. 2009, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls", *British Medical Journal*, vol. 338, b2393.

The EuroQol Group (1990). "EuroQol-a new facility for the measurement of health-related quality of life", Health Policy, vol. 16, no. 3, pp. 199-208.

van Buuren S, Oudshoorn CGM. 2000, "Multivariate Imputation by Chained Equations: MICE V1.0 User's manual", Leiden: TNO Quality of Life. http://www.stefvanbuuren.nl/publications/MICEV1.0ManualTNO000382000.pdf

van Buuren S. 2007, "Multiple imputation of discrete and continuous data by fully conditional specification", *Statistical Methods in Medical Research*, vol. 16, pp. 219-242.

Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE, Walker-Smith JA. 1998, "RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children", *The Lancet*, vol. 351, pp. 637-641.

Walker J, Cassidy J, Sharpe M. 2009a, “The second Symptom Management Research Trial in Oncology (SMaRT Oncology-2): a randomised trial to determine the effectiveness and cost-effectiveness of adding a complex intervention for major depressive disorder to usual care for cancer patients”, *Trials*. Published online, doi:10.1186/1745-6215-10-18.

Walker J, Cassidy J, Sharpe M. 2009b, “The third Symptom Management Research Trial in Oncology (SMaRT Oncology-3): a randomised trial to determine the efficacy of adding a complex intervention for major depressive disorder (Depression Care for People with Lung Cancer) to usual care, compared to usual care alone in patients with lung cancer”, *Trials*. Published online, doi:10.1186/1745-6215-10-92.

Walker J, Holm Hansen C, Martin P, Sawney A, Thekkumpurath P, Beale C, Symeonides S, Wall L, Murray G, Sharpe M. 2012, “Prevalence of depression in adults with cancer: A systematic review”, *Annals of Oncology*. Published online first: doi: 10.1093/annonc/mds575.

Walker J, Postma K, McHugh GS, Rush R, Coyle B, Strong V, Sharpe M. 2007, "Performance of the Hospital Anxiety and Depression Scale as a screening tool for major depressive disorder in cancer patients", J Psychosom.Res., vol. 63, no. 1, pp. 83-91.

Ware, J. H. 2003, “Interpreting Incomplete Data in Studies of Diet and Weight Loss”, *N Engl J Med*, vol. 348, pp. 2136-2137.

White IR, Royston P, Wood AM. 2011, “Multiple imputation using chained equations: Issues and guidance for practice”, *Statistics in Medicine*, vol. 30, pp. 377-399.

Zigmond AS, Snaith RP. 1983, "The hospital anxiety and depression scale", Acta Psychiatrica Scandinavica, vol. 67, pp. 361-370.

APPENDIX A: SUPPLEMENT TO CHAPTER 8

Table A.1. Prevalence and associations of major depression in outpatients with a cancer diagnosis presented separately for five common cancer types.

Lung cancer

	Total n (%)	Major depression Yes n (%)	Major depression No n (%)	Adjusted odds ratio (95% CI)	p-value
Total	4316 (100%)	564 (13%)	3752 (87%)		
Gender					**<.001**
Male	2216 (51%)	238 (11%)	1978 (89%)	1	
Female	2100 (49%)	327 (16%)	1773 (84%)	1.52 (1.24 to 1.87)	
Age group					**<.001**
<50 years	150 (3%)	39 (26%)	111 (74%)	1	
50-59 years	591 (14%)	119 (20%)	472 (80%)	0.76 (0.48 to 1.21)	
60-69 years	1473 (34%)	212 (14%)	1261 (86%)	0.51 (0.33 to 0.79)	
≥70 years	2102 (49%)	194 (9%)	1908 (91%)	0.31 (0.20 to 0.48)	
Time since diagnosis					**0.599**
<1 year	3362 (78%)	447 (13%)	2915 (87%)	1.08 (0.81 to 1.43)	
≥1 year	954 (22%)	117 (12%)	837 (88%)	1	
Recent cancer treatment*					**0.740**
No	2613 (61%)	327 (13%)	2286 (87%)	1	
Yes	1703 (39%)	237 (14%)	1466 (86%)	1.04 (0.83 to 1.30)	
Therapeutic objective					**0.430**
Palliative	3105 (72%)	401 (13%)	2705 (87%)	1	
Curative	1211 (28%)	163 (13%)	1047 (87%)	1.10 (0.87 to 1.37)	
Resident setting					**0.627**
Urban	3598 (83%)	481 (13%)	3117 (87%)	1	
Small town	374 (9%)	50 (13%)	324 (87%)	1.13 (0.78 to 1.64)	
Rural	344 (8%)	33 (10%)	311 (90%)	0.87 (0.57 to 1.34)	
Deprivation SIMD quintile score**					**<.001**
1	1481 (34%)	243 (16%)	1238 (84%)	2.20 (1.50 to 3.21)	
2	1047 (24%)	154 (15%)	893 (85%)	1.95 (1.31 to 2.91)	
3	693 (16%)	81 (12%)	612 (88%)	1.50 (0.97 to 2.33)	
4	536 (12%)	43 (8%)	493 (92%)	0.98 (0.59 to 1.62)	
5	559 (13%)	44 (8%)	515 (92%)	1	

Breast cancer

	Total n (%)	Major depression Yes n (%)	Major depression No n (%)	Adjusted odds ratio (95% CI)	p-value
Total	8461 (100%)	788 (9%)	7673 (91%)		
Gender					**0.980**
Male	28 (0%)	1 (5%)	27 (95%)	1	
Female	8433 (100%)	787 (9%)	7646 (91%)	6.22 (0.00 to 1e64)	
Age group					**<.001**
<50 years	1413 (17%)	205 (14%)	1208 (86%)	1	
50-59 years	2161 (26%)	277 (13%)	1884 (87%)	0.90 (0.73 to 1.11)	
60-69 years	2698 (32%)	223 (8%)	2475 (92%)	0.54 (0.44 to 0.67)	
≥70 years	2189 (26%)	84 (4%)	2104 (96%)	0.24 (0.18 to 0.31)	
Time since diagnosis					**0.145**
<1 year	3467 (41%)	317 (9%)	3150 (91%)	0.77 (0.55 to 1.09)	
≥1 year	4994 (59%)	472 (9%)	4522 (91%)	1	
Recent cancer treatment*					**0.608**
No	5251 (62%)	491 (9%)	4760 (91%)	1	
Yes	3210 (38%)	297 (9%)	2913 (91%)	1.10 (0.77 to 1.57)	
Therapeutic objective					**0.209**
Palliative	553 (7%)	60 (11%)	493 (89%)	1	
Curative	7908 (93%)	728 (9%)	7180 (91%)	0.80 (0.57 to 1.13)	
Resident setting					**0.225**
Urban	6603 (78%)	648 (10%)	5955 (90%)	1	
Small town	820 (10%)	63 (8%)	757 (92%)	0.84 (0.63 to 1.12)	
Rural	1038 (12%)	78 (7%)	960 (93%)	0.82 (0.62 to 1.08)	
Deprivation SIMD quintile score**					**<.001**
1	1458 (17%)	218 (15%)	1240 (85%)	2.89 (2.26 to 3.68)	
2	1559 (18%)	176 (11%)	1383 (89%)	2.18 (1.70 to 2.80)	
3	1553 (18%)	152 (10%)	1401 (90%)	1.84 (1.41 to 2.39)	
4	1627 (19%)	116 (7%)	1511 (93%)	1.34 (1.00 to 1.78)	
5	2264 (27%)	127 (6%)	2137 (94%)	1	

Genitourinary cancer

	Total n (%)	Major depression Yes n (%)	No n (%)	Adjusted odds ratio (95% CI)	p-value
Total	2009 (100%)	113 (6%)	1896 (94%)		
Gender					**0.252**
Male	1926 (96%)	105 (5%)	1821 (95%)	1	
Female	83 (4%)	8 (10%)	75 (90%)	1.68 (0.69 to 4.07)	
Age group					**0.004**
<50 years	164 (8%)	15 (9%)	149 (91%)	1	
50-59 years	200 (10%)	18 (9%)	182 (91%)	0.97 (0.43 to 2.22)	
60-69 years	659 (33%)	47 (7%)	612 (93%)	0.77 (0.38 to 1.55)	
≥70 years	986 (49%)	34 (3%)	952 (97%)	0.36 (0.17 to 0.77)	
Time since diagnosis					**0.672**
<1 year	657 (33%)	44 (7%)	613 (93%)	1.12 (0.66 to 1.92)	
≥1 year	1352 (67%)	69 (5%)	1283 (95%)	1	
Recent cancer treatment*					**0.701**
No	1719 (86%)	92 (5%)	1627 (95%)	1	
Yes	290 (14%)	21 (7%)	269 (93%)	1.14 (0.59 to 2.20)	
Therapeutic objective					**0.687**
Palliative	812 (40%)	46 (6%)	766 (94%)	1	
Curative	1197 (60%)	67 (6%)	1130 (94%)	0.90 (0.54 to 1.51)	
Resident setting					**0.150**
Urban	1505 (75%)	98 (6%)	1407 (94%)	1	
Small town	202 (10%)	8 (4%)	194 (96%)	0.75 (0.31 to 1.78)	
Rural	302 (15%)	7 (2%)	295 (98%)	0.43 (0.18 to 1.00)	
Deprivation SIMD quintile score**					**<.001**
1	352 (18%)	43 (12%)	309 (88%)	10.99 (3.89 to 31.1)	
2	342 (17%)	30 (9%)	312 (91%)	7.69 (2.69 to 22.0)	
3	358 (18%)	16 (4%)	342 (96%)	4.18 (1.37 to 12.8)	
4	415 (21%)	18 (4%)	397 (96%)	4.33 (1.45 to 12.9)	
5	542 (27%)	7 (1%)	535 (99%)	1	

Gynaecological cancer

	Total n (%)	Major depression Yes n (%)	Major depression No n (%)	Adjusted odds ratio (95% CI)	p-value
Total	3010 (100%)	329 (11%)	2681 (89%)		
Age group					**<.001**
<50 years	550 (18%)	105 (19%)	445 (81%)	1	
50-59 years	607 (20%)	98 (16%)	509 (84%)	0.88 (0.64 to 1.21)	
60-69 years	892 (30%)	74 (8%)	818 (92%)	0.41 (0.29 to 0.57)	
≥70 years	961 (32%)	51 (5%)	910 (95%)	0.25 (0.17 to 0.36)	
Time since diagnosis					**0.036**
<1 year	1529 (51%)	179 (12%)	1350 (88%)	1.47 (1.03 to 2.11)	
≥1 year	1481 (49%)	151 (10%)	1330 (90%)	1	
Recent cancer treatment*					**0.052**
No	1842 (61%)	204 (11%)	1638 (89%)	1	
Yes	1168 (39%)	126 (11%)	1043 (89%)	0.70 (0.48 to 1.00)	
Therapeutic objective					**0.984**
Palliative	653 (22%)	66 (10%)	587 (90%)	1	
Curative	2357 (78%)	263 (11%)	2094 (89%)	1.00 (0.72 to 1.41)	
Resident setting					**0.180**
Urban	2275 (76%)	265 (12%)	2010 (88%)	1	
Small town	330 (11%)	24 (7%)	306 (93%)	0.64 (0.39 to 1.05)	
Rural	405 (13%)	40 (10%)	365 (90%)	0.99 (0.68 to 1.46)	
Deprivation SIMD quintile score**					**<.001**
1	633 (21%)	115 (18%)	518 (82%)	2.63 (1.77 to 3.92)	
2	649 (22%)	71 (11%)	578 (89%)	1.44 (0.94 to 2.20)	
3	572 (19%)	50 (9%)	522 (91%)	1.19 (0.76 to 1.87)	
4	566 (19%)	52 (9%)	514 (91%)	1.20 (0.76 to 1.90)	
5	590 (19%)	42 (7%)	548 (93%)	1	

Lower gastrointestinal cancer

	Total n (%)	Major depression Yes n (%)	Major depression No n (%)	Adjusted odds ratio (95% CI)	p-value
Total	3355 (100%)	236 (7%)	3119 (93%)		
Gender					**0.031**
Male	1869 (56%)	112 (6%)	1757 (94%)	1	
Female	1486 (44%)	124 (8%)	1362 (92%)	1.38 (1.03 to 1.86)	
Age group					**<.001**
<50 years	244 (7%)	37 (15%)	207 (85%)	1	
50-59 years	545 (16%)	75 (14%)	470 (86%)	0.97 (0.61 to 1.53)	
60-69 years	1098 (33%)	70 (6%)	1028 (94%)	0.43 (0.27 to 0.68)	
≥70 years	1468 (44%)	54 (4%)	1414 (96%)	0.24 (0.14 to 0.39)	
Time since diagnosis					**0.395**
<1 year	1679 (50%)	124 (7%)	1555 (93%)	1.19 (0.80 to 1.76)	
≥1 year	1676 (50%)	112 (7%)	1564 (93%)	1	
Recent cancer treatment*					**0.241**
No	2217 (66%)	158 (7%)	2059 (93%)	1	
Yes	1138 (34%)	78 (7%)	1060 (93%)	0.78 (0.52 to 1.18)	
Therapeutic objective					**0.645**
Palliative	768 (23%)	58 (7%)	710 (93%)	1	
Curative	2587 (77%)	178 (7%)	2409 (93%)	0.92 (0.64 to 1.32)	
Resident setting					**0.163**
Urban	2708 (81%)	196 (7%)	2512 (93%)	1	
Small town	275 (8%)	24 (9%)	251 (91%)	1.45 (0.89 to 2.35)	
Rural	372 (11%)	16 (4%)	356 (96%)	0.74 (0.41 to 1.34)	
Deprivation SIMD quintile score**					**<.001**
1	648 (19%)	82 (13%)	566 (87%)	3.21 (2.02 to 5.08)	
2	662 (20%)	51 (8%)	611 (92%)	1.90 (1.17 to 3.10)	
3	605 (18%)	43 (7%)	562 (93%)	1.82 (1.08 to 3.05)	
4	587 (17%)	26 (4%)	561 (96%)	1.09 (0.61 to 1.95)	
5	853 (25%)	33 (4%)	820 (96%)	1	

Notes: Missing data for the depression response, recent treatment and therapeutic objective were handled using multiple imputation with the reported frequencies averaged over the m=100 imputed datasets. *Any of chemotherapy, radiotherapy or surgery started in the preceding six months. **Scottish Index of Multiple Deprivation quintile score: 1=most deprived, 5=least deprived. Odds ratios are conditional on all other variables in the model.

APPENDIX B: PUBLISHED WORK

'Reprinted with permission'

Psychological Medicine, Page 1 of 8. © Cambridge University Press 2013
doi:10.1017/S0033291712002930

ORIGINAL ARTICLE

Screening medical patients for distress and depression: does measurement in the clinic prior to the consultation overestimate distress measured at home?

C. H. Hansen[1], J. Walker[2], P. Thekkumpurath[1], A. Kleiboer[1], C. Beale[1], A. Sawhney[1], G. Murray[3] and M. Sharpe[2]*

[1] *Psychological Medicine Research, School of Molecular and Clinical Medicine, University of Edinburgh, UK*
[2] *Psychological Medicine Research, Department of Psychiatry, University of Oxford, UK*
[3] *Centre for Population Health Sciences, University of Edinburgh, UK*

Background. Medical patients are often screened for distress in the clinic using a questionnaire such as the Hospital Anxiety and Depression Scale (HADS) while awaiting their consultation. However, might the context of the clinic artificially inflate the distress score? To address this question we aimed to determine whether those who scored high on the HADS in the clinic remained high scorers when reassessed later at home.

Method. We analysed data collected by a distress and depression screening service for cancer out-patients. All patients had completed the HADS in the clinic (on computer or on paper) prior to their consultation. For a period, patients with a high score (total of $\geqslant$15) also completed the HADS again at home (over the telephone) 1 week later. We used these data to determine what proportion remained high scorers and the mean change in their scores. We estimated the effect of 'regression to the mean' on the observed change.

Results. Of the 218 high scorers in the clinic, most [158 (72.5%), 95% confidence interval (CI) 66.6–78.4] scored high at reassessment. The mean fall in the HADS total score was 1.74 (95% CI 1.09–2.39), much of which could be attributed to the estimated change over time (regression to the mean) rather than the context.

Conclusions. Pre-consultation distress screening in clinic is widely used. Reassuringly, it only modestly overestimates distress measured later at home and consequently would result in a small proportion of unnecessary further assessments. We conclude it is a reasonable and convenient strategy.

Received 21 April 2012; Revised 20 September 2012; Accepted 23 November 2012

Key words: Cancer, depression, distress, screening, test–retest.

Introduction

There is increasing awareness of the importance of subjective measures including quality of life in medical care. Such measure are often referred to as patient-reported outcomes or 'PROs' (Greenhalgh, 2009). Emotional distress and depression are important PROs that have a major effect on quality of life (Moussavi *et al.* 2007). Consequently, it has been recommended that medical patients, such as those with cancer (Carlson *et al.* 2012), are screened for emotional distress and depression (Pignone *et al.* 2002; NICE, 2009), but only if there are facilities to provide treatment for identified cases (USPSTF, 2009). Despite an extensive literature on such screening (Carlson *et al.* 2012), there is limited information on the practicalities of carrying it out, an important aspect of which is when and where to administer the screening measures.

The most convenient and widely used strategy is to administer a questionnaire, such as the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983), in the medical clinic, taking advantage of the time patients spend waiting to go into their consultation. The patient's questionnaire score is then used to determine whether they have a significant level of distress that requires attention and whether they need a further assessment to determine whether they have a depressive disorder.

However, there is a potential problem with this strategy; measuring distress in the clinic prior to the consultation might result in a transient inflation

* Address for correspondence: Professor M. Sharpe, Psychological Medicine Research, Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford OX3 7JX, UK.
(Email: michael.sharpe@psych.ox.ac.uk)

of the score because of the clinical context and the anticipation of the consultation. This phenomenon would be similar to that referred to as the 'white-coat effect' in the measurement of blood pressure (Gerin *et al.* 2006). If such inflation were to occur it would result in false positives in the identification of patients suffering from significant distress and would lead to more patients than necessary being given assessment interviews for depression. Such an effect would therefore be important in increasing both inconvenience to patients and the costs to clinical services.

As far as we are aware, although there are studies of the test–retest reliability of measures of quality of life and distress (Hjermstad *et al.* 1995; Bakker *et al.* 2009), the course of distress over a series of cancer consultations (van Dooren *et al.* 2005) and of the influence of the content of the consultation on distress (van Dulmen *et al.* 1995), this particular question has not been specifically addressed in the published literature.

We therefore aimed to find out whether oncology patients who were high scorers on the HADS questionnaire, completed while waiting for their cancer consultation in clinic, remained high scorers when completing a repeat HADS questionnaire a week later at home. Specifically, we aimed to determine: (*a*) what proportion of the patients who scored high (total score of $\geqslant$15) on the HADS prior to their consultation still had a high score when reassessed at home 1 week later; and (*b*) how much the mean HADS score had changed between these two occasions and how much any fall could be accounted for by regression to the mean.

Method

To address the research question we analysed data that had been routinely collected by an established distress and depression screening service operating in multiple cancer out-patient clinics in Scotland, UK.

Routine screening procedure

The screening service was in operation in numerous clinics, each specializing in one of a variety of cancer types including breast, colorectal, gynaecological, lung and genito-urinary. All patients attending the clinics were asked to complete the HADS on touch-screen computers (or, where computers were not available, on paper) prior to their medical consultation. The results of screening were given to their cancer clinician at the time of the consultation. In addition, all patients who had scored high on the HADS in clinic were telephoned at home, approximately 1 week later, and assessed for depression using the major depression component of the Structured Clinical Interview for DSM-IV (SCID; First *et al.* 1999).

Collection of repeat HADS scores

As part of routine clinical service data collection during March and April 2009, patients who had scored $\geqslant$15 on the HADS in the clinic were asked to complete the HADS again at home over the telephone, immediately before they were given the routine interview to assess them for depression. We analysed these clinical data to address the research question.

Ethical approval

We obtained ethical approval from the local Research Ethics Committee to use the data in this way and also obtained each patient's permission to use their anonymized clinical data for research.

Measure

The HADS is the most extensively studied distress scale in cancer patients and is very widely used as a first stage in screening medical patients for depression (Vodermaier *et al.* 2009). The HADS asks patients how they have been feeling over the past 2 weeks. It has 14 items: seven on each of the anxiety and depression subscales. Each item is rated from 0 to 3, resulting in a total HADS score between 0 and 42, with higher scores indicating more severe symptoms (Zigmond & Snaith, 1983) A recent review concluded that the HADS was an effective measure of emotional distress but that the subscales were unable to differentiate consistently between anxiety and depression (Cosco *et al.* 2012). A total HADS score of $\geqslant$15 has been reported to be optimal to identify cancer patients likely to have major depression on further assessment (Walker *et al.* 2007).

Analysis

We analysed these data to determine whether patients with high HADS scores measured in the clinic prior to their consultation still had high scores when measured later at home. We therefore calculated the proportion of patients who still had a high score ($\geqslant$15) when the HADS was repeated at home. We also determined the mean change in the total HADS score between clinic and home.

Individual patient distress scores vary over time. Patients scoring high or low are likely to score closer to the mean score of all assessed patients on later reassessment, a phenomenon known as 'regression to the mean'. If all patients who completed a first HADS also completed a second HADS, we would expect the effect of these variations on the mean score of the

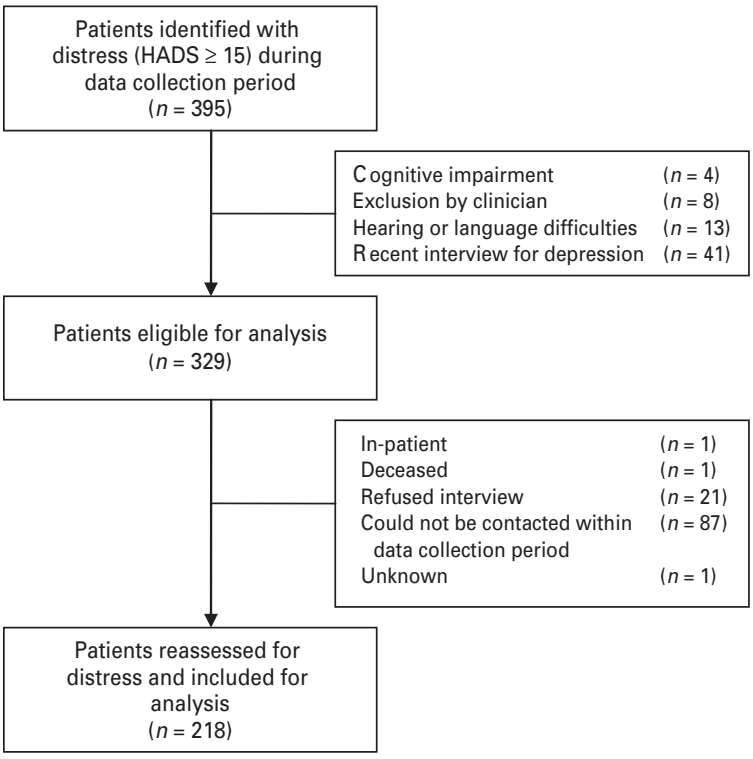


Fig. 1. Derivation of patient sample. The patients initially identified with distress (HADS ⩾15) were screened during the period from 25 February 2009 to 31 March 2009.

whole group to even out. However, as we only had follow-up data the subsample of initial high scorers we would expect the average of the reassessed scores in this subsample to be lower because of this 'regression to the mean' effect (Barnett *et al.* 2005). Therefore, to isolate the effect of the clinic from this phenomenon we estimated the size of the anticipated regression to the mean. This involved using more than 5000 HADS scores that had been collected by the screening service in similar clinics from 2007 to 2010 to obtain details of the overall distribution of HADS scores in this population. These details included the variance and covariance of repeated scores. The technical details of this approach are provided in the Appendix and described elsewhere (Das & Mulder, 1983). Finally, we conducted an exploratory analysis describing and comparing the changes in the HADS anxiety and depression subscales to determine whether these differed in the amount they changed.

Results

The service had offered screening to all patients attending the cancer clinics except for a small number (<5%) who were unable to complete questionnaires because they were too unwell or had severe cognitive or communication problems. A further 10% of patients were missed by the service, mainly because they were taken straight to their consultation before being screened, and an additional 7% refused to participate in screening.

A total of 1691 patients were screened in clinic during the period from which the data analysed were derived. Of these, 395 scored high on the HADS in clinic and 329 were listed for further assessment at home (the remainder were not listed for a variety of reasons including a recent depression assessment, cognitive or communication problems or exclusion by their clinician, usually because they were considered to be too ill). Repeat HADS were not available on 111 of these patients for several reasons, but mainly because they were not contacted by the screening service within the 1-month time window used for the analysis. The final patient sample is shown in Fig. 1. A total of 218 patients were given a repeat HADS at home by the screening service during the data collection period. This is the sample analysed.

In the analysed sample, 159 (73%) patients were female and the median age was 61 years [interquartile range (IQR) 53–70 years]. Almost all of the patients were attending follow-up appointments. The median interval between the clinic and repeat HADS assessments was 6 days (IQR 5–8 days). The 111 patients who did not have a repeat HADS at home had similar

Table 1. *Characteristics of the analysed sample compared with those in the group of eligible patients not included*

	Eligible patients included for analysis ($n=218$)	Eligible patients not included for analysis ($n=111$)	p value[a]
Age (years)			0.954
Mean (S.D.)	61.4 (11.5)	61.3 (12.2)	
Median (range)	61.4 (25.3–87.7)	62.5 (28.7 to 89.8)	
Age categories, *n* (%)			0.736
⩽50 years	38 (17)	23 (21)	
51–60 years	67 (31)	28 (25)	
61–70 years	66 (30)	35 (32)	
⩾71 years	47 (22)	25 (23)	
Gender, *n* (%)			0.396
Male	59 (27)	35 (32)	
Female	159 (73)	76 (68)	
Cancer clinic type, *n* (%)			0.383
Breast	83 (38)	39 (35)	
Gynaecology	33 (15)	15 (14)	
Lung	47 (22)	24 (22)	
Colorectal	18 (8)	8 (7)	
Urology	14 (6)	12 (11)	
Gastrointestinal	15 (7)	4 (4)	
Other	8 (4)	9 (8)	
Appointment type[b], *n* (%)			0.025
First appointment	30 (14)	6 (6)	
Return appointment	183 (86)	100 (94)	
Poor prognosis[c], *n* (%)			<0.001
Yes	19 (9)	25 (23)	
No	195 (91)	84 (77)	
HADS scores			0.356
Mean (S.D.)	20.1 (4.7)	20.6 (4.8)	
Median (range)	19 (15–37)	19 (15–34)	
HADS score categories, *n* (%)			0.604
15–19	115 (53)	59 (53)	
20–24	66 (30)	29 (26)	
⩾ 25	37 (17)	23 (21)	

HADS, Hospital Anxiety and Depression Scale; S.D., standard deviation.

[a] Age in years and HADS scores were compared using the Wilcoxon rank sum test. All other p values were from χ^2 tests.

[b] Appointment type was unknown for 10 patients.

[c] Poor prognosis was defined for lung (non-lung) cancer patients as a life expectancy of <3 (12) months. Prognosis was unknown for six patients.

distributions of sex, age and clinic HADS scores and attended similar types of cancer clinics. However, there were more new and good prognosis patients included in the sample reassessed. The patients' characteristics and the comparison of those with and without a HADS rated at home are shown in Table 1.

Figure 2 shows the distributions of HADS scores when patients were (*a*) assessed in clinic and (*b*) re-assessed at home. Fig. 3 shows the change in HADS scores for each individual patient. As a result of the large variance in the HADS scores, there was also considerable variability in the change scores between the two assessments despite a high intra-class correlation between repeated measurements (ICC = 0.83).

Almost three-quarters (72.5%; 158/218) of the initial high-scoring patients were still high scorers at reassessment [95% confidence interval (CI) 66.6–78.4]. The mean change in total HADS score was a reduction of 1.74 points (95% CI 1.09–2.39).

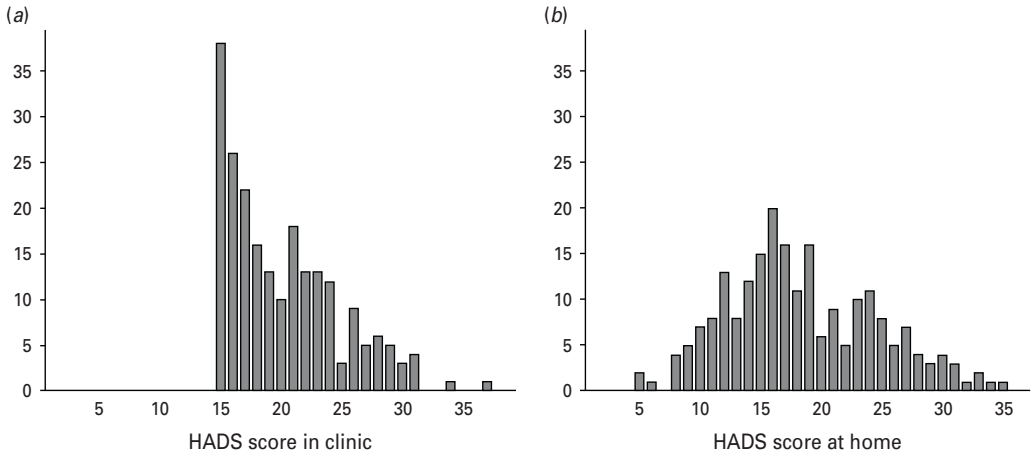


Fig. 2. Hospital Anxiety and Depression Scale (HADS) scores of patients ($n = 218$) in the study sample (*a*) when assessed in clinic and (*b*) when reassessed at home.

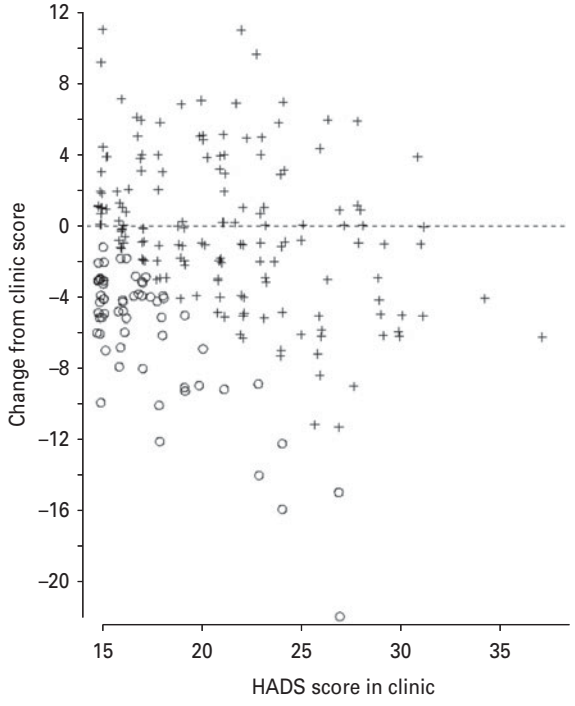


Fig. 3. Change in Hospital Anxiety and Depression Scale (HADS) total score from clinic to follow-up at home plotted against initial HADS score in clinic. Circles indicate patients whose reassessment score fell below 15. Patients plotted above the dashed line had a higher HADS score on reassessment whereas those below the line had a lower score. A degree of 'jitter' was applied to separate out overlapping data points.

Our estimate of the regression-to-mean effect was an average reduction of 1.21 points (95% CI 1.02–1.43). Hence regression to the mean potentially accounts for the majority of this observed fall in mean score, meaning that the effect of measuring in clinic was very small. The exploratory analysis of changes in HADS subscales found a mean reduction in the anxiety subscale of 1.26 points (95% CI 0.84–1.67) and in the depression subscale of 0.48 points (95% CI 0.12–0.85). The difference between the scales in the reduction in scores was statistically significant ($p < 0.001$).

Discussion

We had hypothesized that patients' HADS scores might be transiently inflated when measured in the clinic prior to the consultation because of the potentially stressful clinical surroundings and anticipation of the upcoming appointment. If that were the case it would question the utility of this widely used strategy for screening for distress and depression in medical clinics. We found that the majority of the patients who scored high on the HADS in clinic prior to their cancer consultation (72.5%) were still high scorers when reassessed at home a week later. That also means that 27.5% of patients who had scored high in the clinic were no longer high scorers when reassessed later at home. However, further analysis indicates that despite large variability at the individual patient level, the mean HADS total score in the sample fell by only 1.74 points between the two assessments, most of which could reasonably be attributed to the natural tendency for individuals who score high on an initial measurement to score lower on later reassessment (regression to the mean), independent of the setting in which the measurement was made. Our hypothesis was therefore not supported and measuring distress in the clinic prior to the consultation is a reasonable strategy to adopt.

There was considerable individual variability in the size of change scores between the two assessments

despite a high intra-class correlation between repeated measurements from the same patient. This was due to large overall variance in the scores, a property common to measures of psychological distress. It is unclear whether this variation is due to a large random error in the measurements or a reflection of actual fluctuations in the severity of distress over time. Nonetheless, our sample of 218 patients was sufficiently large to estimate the mean change for the sample with reasonable accuracy.

It is notable that, whereas the screening service used the total score in the HADS to define significant distress, the fall in score was slightly larger on the anxiety subscale. This may be because the consultation has a greater transient effect on anxiety than on depression. It may also imply that scales that measure only depressive symptoms are even less subject to a clinic effect.

We are not aware of any studies that have directly addressed the question we have posed. We identified a test–retest reliability study of the European Organization for Research and Treatment of Cancer (EORTC) quality of life measures, which include emotional functioning, that compared questionnaire scores administered to 270 patients attending routine post-treatment follow-up visits to cancer clinics with their score at home 4 days later and found generally good agreement (Hjermstad *et al.* 1995). Other studies that have administered repeated psychological assessment have examined distress trajectories over longer periods of time (Hinnen *et al.* 2008) or before and after consultations (van Dooren *et al.* 2005) but we found none that directly addressed the possible effect of the clinical context on the measurement score.

There were limitations to this study. First, we analysed data collected by a routine screening service operating in cancer clinics; the findings may not therefore generalize to other clinical settings. Second, the service administered a second HADS only to patients who had scored high in clinic. This meant that our observed HADS scores obtained at home underestimated the true proportion of patients who would have scored high had all patients been reassessed, as it would be likely that some of the patients who scored low in clinic would have scored high on the second occasion. This limitation was addressed by estimating the regression to the mean. Third, there were missing data from patients who could not be contacted during the limited time window in which repeat HADS were administered. However, the characteristics of patients on whom we had analysable data and those on whom we did not were mostly similar; systematic bias is therefore unlikely. Fourth, there may be limits to the intrinsic test–retest reliability of the HADS (as opposed to real changes in symptoms) but this is unlikely to be large over this time period, or to represent a systematic bias. Fifth, patients completed the HADS on a touch-screen computer or on paper in the clinic, but the follow-up assessment was carried out by reading out the scale over the telephone. It is possible that administering the HADS over the telephone causes patients to score differently. Previous studies have found good agreement between self-completed and verbally completed distress screening questionnaires, with a tendency for the latter to record a lower score (Pinto-Meza *et al.* 2005; Cheung *et al.* 2006). Such a bias, if present, would reduce further the observed fall in HADS score attributable to the effect of measurement in the clinic. Future studies could use the same mode of administration to avoid this issue. Sixth, the content of the consultation and its meaning for the patient, whether positive or negative, might have accounted for some of the changes in scores and we were not able to assess this. However, most of the consultations were for follow-up and not for the communication of new diagnoses. The effect of consultation type could be addressed in future studies. Seventh, because the results of the screening were given to the clinician before the consultation, it is possible that they might have taken action to address the distress, for example by referring the patient for psychological treatment. This is, however, very unlikely to have occurred within 1 week. Finally, although the average change in scores was small, the intra-patient variability was high, with some patients scoring very differently on reassessment. It is possible, therefore, that a minority of patients are affected considerably by the clinic setting. Consequently, we cannot rule out the possibility of an important 'clinic effect' for some individuals.

Conclusions

In conclusion, most patients who scored high on the HADS administered in clinic prior to their medical consultation remained high scorers when reassessed at home a week later. There was only a small reduction in mean score, most of which could be attributed to regression to the mean. Therefore, the widely used strategy of asking patients to complete a screening questionnaire for distress while they wait for their clinic appointment is a reasonable method of identifying those who have significant distress and also a useful first step in identifying those who require an interview for the assessment of possible depressive disorder. The increasing use of telephones and the internet provides opportunities to screen patients away from the clinic, thereby potentially avoiding the issue of clinic context. However, the pre-consultation waiting time has long provided an opportunity to

undertake such clinic-based screening, and is likely to continue to do so in the future.

Appendix

Estimating the regression-to-the-mean effect

As only patients with an initial high score were followed up, the scores on reassessment were subject to regression to the mean. We estimated the average drop in scores caused by this effect as follows.

Suppose that a patient's HADS score, H, is the sum of their (constant) true underlying score, S, and an independent error term, e, where S is distributed according to some arbitrary density function with variance σ_s^2 and the errors are normally distributed with mean 0 and variance σ_e^2. The total variance is then $\mathrm{Var}(H) = \sigma_t^2 = \sigma_s^2 + \sigma_e^2$ and $\rho = \sigma_s^2/\sigma_t^2$ is the intra-patient correlation between repeated scores from the same individual.

We wanted to estimate the expected difference between a pair of repeated HADS scores, H_1 and H_2, conditional on H_1 being $\geqslant 15$. That is, we wanted to estimate $E[H_1 - H_2 | H_1 \geqslant 15]$.

For a continuous H it can be shown that

$$E[H_1 - H_2 | H_1 > h_c] = (1-\rho)\sigma_t^2 \frac{g(h_c)}{1-G(h_c)},$$

where $g(h_c)$ is the probability density function for H evaluated at h_c, and $G(h_c)$ is the corresponding cumulative distribution function (Das & Mulder, 1983). From the large sample of scores collected by the screening service in similar clinics from 2007 to 2010, we obtained empirical estimates of $g(h_c)$ and $G(h_c)$. Using data from the 5215 patients who had HADS scores measured in subsequent clinic visits during this period, we estimated σ_t^2 and ρ as the correlation of scores obtained 1 week apart. We did this by modelling the covariance matrix of repeated scores in a linear regression with random intercept and exponential covariance structure to account for a decreasing correlation over time. A 95% quantile-based CI for the regression-to-mean estimate was derived through bootstrapping.

Although technically a discrete scale, we applied the HADS (range 0–42) with the above result, introducing a continuity correction by evaluating $g(.)$ and $G(.)$ at $h_c = 14.5$ by approximating a theoretical continuous curve. The approach was verified through simulation studies and sensitivity analysis.

Acknowledgements

We thank the clinical service and patients who provided the data. This work was funded as part of a Programme Grant from Cancer Research UK (CRUK), ref. C5547/A7375.

Declaration of Interest

None.

References

Bakker IM, Terluin B, van Marwijk HW, van Mechelen W, Stalman WA (2009). Test-retest reliability of the PRIME-MD: limitations in diagnosing mental disorders in primary care. *European Journal of Public Health* **19**, 303–307.

Barnett AG, van der Pols JC, Dobson AJ (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology* **34**, 215–220.

Carlson LE, Waller A, Mitchell AJ (2012). Screening for distress and unmet needs in patients with cancer: review and recommendations. *Journal of Clinical Oncology* **30**, 1160–1177.

Cheung YB, Goh C, Thumboo J, Khoo KS, Wee J (2006). Quality of life scores differed according to mode of administration in a review of three major oncology questionnaires. *Journal of Clinical Epidemiology* **59**, 185–191.

Cosco TD, Doyle F, Ward M, McGee H (2012). Latent structure of the Hospital Anxiety and Depression Scale: a 10-year systematic review. *Journal of Psychosomatic Research* **72**, 180–184.

Das P, Mulder PGH (1983). Regression to the mode. *Statistica Neerlandica* **37**, 15–20.

First MB, Spitzer RL, Gibbon M, Williams JBW (1999). *Structured Clinical Interview for DSM-IV Axis I Disorders*. Biometrics Research Department, New York State Psychiatric Institute: New York.

Gerin W, Ogedegbe G, Schwartz JE, Chaplin WF, Goyal T, Clemow L, Davidson KW, Burg M, Lipsky S, Kentor R, Jhalani J, Shimbo D, Pickering TG (2006). Assessment of the white-coat effect. *Journal of Hypertension* **24**, 67–74.

Greenhalgh J (2009). The applications of PROs in clinical practice: what are they, do they work, and why? *Quality of Life Research* **18**, 115–123.

Hinnen C, Ranchor AV, Sanderman R, Snijders TA, Hagedoorn M, Coyne JC (2008). Course of distress in breast cancer patients, their partners, and matched control couples. *Annals of Behavioral Medicine* **36**, 141–148.

Hjermstad MJ, Fossa SD, Bjordal K, Kaasa S (1995). Test/retest study of the European Organization for Research and Treatment of Cancer Core Quality-of-Life Questionnaire. *Journal of Clinical Oncology* **13**, 1249–1254.

Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* **370**, 851–858.

NICE (2009). *Depression in Adults with a Chronic Physical Health Problem: Treatment and Management*. National Institute for Health and Clinical Excellence: London.

Pignone MP, Gaynes BN, Rushton JL, Burchell CM, Orleans CT, Mulrow CD, Lohr KN (2002). Screening for depression in adults: a summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine* **136**, 765–776.

Pinto-Meza A, Serrano-Blanco A, Penarrubia MT, Blanco E, Haro JM (2005). Assessing depression in primary care with the PHQ-9: can it be carried out over the telephone? *Journal of General Internal Medicine* **20**, 738–742.

USPSTF (2009). Screening for depression in adults: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine* **151**, 784–792.

van Dooren S, Seynaeve C, Rijnsburger AJ, Duivenvoorden HJ, Essink-Bot ML, Tilanus-Linthorst MM, Klijn JG, de Koning HJ, Tibben A (2005). Exploring the course of psychological distress around two successive control visits in women at hereditary risk of breast cancer. *European Journal of Cancer* **41**, 1416–1425.

van Dulmen AM, Fennis JF, Mokkink HG, van der Velden HG, Bleijenberg G (1995). Doctor-dependent changes in complaint-related cognitions and anxiety during medical consultations in functional abdominal complaints. *Psychological Medicine* **25**, 1011–1018.

Vodermaier A, Linden W, Siu C (2009). Screening for emotional distress in cancer patients: a systematic review of assessment instruments. *Journal of the National Cancer Institute* **101**, 1464–1488.

Walker J, Postma K, McHugh GS, Rush R, Coyle B, Strong V, Sharpe M (2007). Performance of the Hospital Anxiety and Depression Scale as a screening tool for major depressive disorder in cancer patients. *Journal of Psychosomatic Research* **63**, 83–91.

Zigmond AS, Snaith RP (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica* **67**, 361–370.